Library Progress International *Print version ISSN 0970 1052*
Vol.43 No.1, January-June 2023: P.301-309 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

Ordinal Data Classification using Support Vector machines with Imbalance Data

Ashoka Wilson Dsouza*, Ismail B

Ashoka Wilson Dsouza*, Research scholar, Department of PG Studies and Research in Statistics, Mangalore University, Konaje, India.
Email: ashokdesouza@gmail.com
Dr Ismail B, Professor of Statistics, Department of Statistics, Yenepoya (Deemed to be University), Deralakatte, Mangalore, India, Email: prof.ismailb@gmail.com

How to cite this article: Ismail B, Ashoka Wilson Dsouza (2023). Ordinal Data Classification using Support Vector machines with Imbalance Data. *Library Progress International,* 42(1), 301-309

Abstract: *Ordinal data classification presents a complex challenge of training a model to correctly categorize observations within ranked data. However, real-world datasets utilized in ordinal classification frequently exhibit imbalanced class distributions, which pose a persistent obstacle in accurately classifying ranked data. The class imbalance issue often results in a bias toward majority classes within most classification models, leading to reduced accuracy and precision for minority classes. Additionally, successful implementations of classification models in finance sector with imbalanced classes remain limited.. Hence, this research paper introduces an innovative approach involving a hybrid class balancing technique followed by the utilization of Support Vector Machines (SVM) as the classifier for classifying mutual fund rating that are ordinal in nature. The study comprehensively compares the proposed hybrid SVM model and ordinal logistic regression in imbalanced data. Furthermore, the research extends its application to predicting mutual fund ratings and other relevant ordinal class data scenarios. Through empirical investigations conducted on both artificial and real-world datasets, including an application in Mutual fund rating analysis, this research establishes the efficacy and practical utility of the proposed approach.*

Based on an empirical experiment, we assess the effectiveness of SVMs in identifying mutual fund ratings, along with implementation resampling methods typically utilized to tackle class imbalances.

Keywords: *Mutual funds classification, Unsupervised learning, weighted SVM, Imbalanced data, Ordinal data, Resampling methods*

INTRODUCTION

Ordinal data classification, also known as Ranking Learning, constitutes a significant supervised problem encompassing both the classification and regression domains. Within the realm of ordinal rank regression, the primary objective is to allocate data into finite collection of ordered categories. For instance, consider an assessment of students' performance using grades A, to E, arranged in a specific order (A > B > C > D > E) [9]. The field of ordinal regression, or ordinal classification, has garnered growing interest primarily due to its relevance in learning-to-rank and reviewing product rating applications.

Though there is significant research towards classifying ordinal data with balanced classes, the imbalanced data with unequal sample size possess a major challenge for the classifier to rightly predict the ordinal class. Imbalanced ordinal data is a common occurrence in real-

world scenarios, requiring the application of specialized solutions and techniques to consider the order of data and rectify class distribution disparities. Ordinal regression stands apart from traditional classification, primarily owing to the intrinsic order and hierarchy of its categories.

SUPPORT VECTOR MACHINES(SVM) For Classification Modelling: SVM is a supervised learning method which can be used to handle both regression and classification tasks (Vapnik, 1998). By utilizing various algorithms and kernels, SVM enables effective data analysis for both classification and regression purposes (Cortes and Vapnik, 1995). One such algorithm, Support Vector Classification (SVC), is designed to determine the optimal separating hyperplane, particularly for linearly separable cases (Burbidge and Buxton, 2001). However, in scenarios where perfect separation between classes is not possible, SVM kernel methods are employed. Commonly used kernels in SVM include polynomial, quadratic, and radial basis functions.

Let's look at an case of a two-class binary classification problem represented as {($_1$, $_1$),(x_2, y_2), . . . ,(x_k, y_k)}, where $x_i \in R^n$ denotes data points in n-dimensional space and $y_i \in \{-1, 1\}$ denotes class to which the data point belongs, for i = 1, . . . , k. Finding the best separating hyperplane to efficiently separates these data points into respective classes is the goal of the SVM classifier learning technique. To improve class segregation, the data points are first transformed using a non-linear mapping function Φ into a feature space of higher dimension, which is represented by,

$$w \cdot \Phi(x) + b = 0 \qquad (1)$$

In this case, the weight vector is perpendicular to the hyperplane is denoted by . When the dataset is linear and completely separable, the following optimization problem centered on maximizing the margin may be utilized to identify the hyperplane that obtains the highest margin (thereby improving generalization capacity).:

$$\min(\tfrac{1}{2} \quad . \ )$$

$$\text{s.t} \quad y_i(w \cdot \Phi(x_i) + b) \geq 1$$

$$\text{i=1,….,k} \qquad (2)$$

Class imbalance impact on Support Vector Machines (SVM):

SVM classifier determines the ideal α_i for each datapoint xi to maximize the margin β between the hyperplane and the closest data instances to it, given a kernel function K and a set of labelled instances $x_{Train} = (x, y)^n$ _. For a new instance x of test data, a class prediction is made using:

$$\text{sign}\{f(x) = \sum^n_{=} \quad (\ , \ ) + \ \} \qquad (3)$$

where b is the threshold. The primal Lagrangian for 1-norm soft SVM minimize is given below

$$L = \frac{\|w\|^2}{2} + C\sum_{=1}^{n} \xi - \sum_{=1}^{n} \alpha\,[y\,(w.x + b) - 1 + \xi] - \sum_{1}^{n} r\xi \quad (4)$$

Where $\alpha_i \geq 0$ and $r_i \geq 0$. Here C the penalty constant represents the trade-off between the margin and empirical error ξ_i. α_i must satisfy the below conditions to satisfy Karush-Kuhn-Tucker test

$$0 \leq \alpha_i \leq C \text{ and } \sum_{=}^{n} = 0 \quad (5)$$

When an SVM model classifier is trained on a dataset with imbalance, the resultant models often demonstrate a bias towards the majority class. Consequently, this bias can result in reduced performance in accurately predicting instances from the minority class. Hence to address this issue in SVMs, researchers have introduced a range of data preprocessing, class balancing/resampling and algorithmic techniques.

Resampling methods:

By giving the classifier only a selected fraction of the available data, resampling aims to alleviate the problem of one class being overrepresented. The basic idea is that until a certain ratio of majority to minority examples is reached, either instances from the majority class population are randomly removed (under-sampling) or instances from the minority class are randomly duplicated (oversampling). Oversampling may have the disadvantage of making the minority class's decision region too particular, which might result in overfitting problems. SVMs with unbalanced datasets have been successfully trained using resampling strategies in a variety of fields. The following: 10, 11, 12, 13, 14, 15, 16

LITERATURE REVIEW

The history of ordinal regression research, which is a case under limited dependent model, can be dated back to 1980s when ordinal regression methods were first introduced. Ordinal regression techniques were also developed in the 1990s because of machine learning research. It has attracted a lot of attention lately due to its potential uses in a variety of data-intensive fields, including protein ranking [8] in bioinformatics and ranking or ratings [24] in the field of finance. Numerous machine learning classifiers have emerged or been adapted to tackle the ordinal regression problem [33]. These include neural networks employing gradient descent [7], [5], Gaussian processes [10], [9], [37], support vector machines [21], [22], [24], [38], [11], [2], [12]], regression trees [25], Naive Bayes [42], and binary classification approaches [16], [26]. These methods aim to decompose the original ordinal regression problem into a series of binary classifications.

The support vector method, known for its effectiveness as a large-margin classifier [38], [11], is adapted for ordinal regression tasks. It works by determining K-1 thresholds that partition the data into K ordered categories. An important feature of this approach is that the complexity of the optimization problem increases linearly with the number of training examples

In this study, we use SVM classification model for imbalanced data and also apply it to for mutual fund rating classification example. A SVM model is developed to avoid the

majority ratings and improve the performance of the minority class rating by employing an imbalanced dataset collected from Indian mutual funds and also tested on other datasets. Additionally, this research seeks to examine.

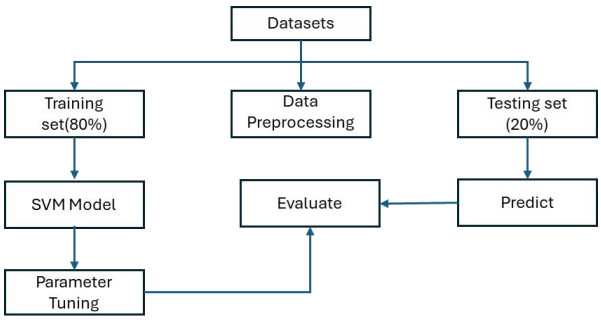
1) How can we address the issue of skewed class ratings when dealing with multiple classes?

2) The existing methods to address the imbalanced data in ordinal datasets.

3) How accurate is the SVM with both imbalanced and balanced datasets using linear and RBF kernels

Support Vector Machines (SVMs) have emerged as a highly favoured machine learning method, proving effective in solving a multitude of real-world classification challenges across a broad class of fields [1- 6]. A SVM classifier model works as a discriminative classifier, formally characterized by its ability to establish a differentiating hyperplane. In simplified terms, in supervised learning scenarios where labelled training data is provided, Support Vector Machines (SVMs) identifies and generates an optimal hyperplane. This hyperplane acts as a tool to accurately classify new instances into separate categories. Studies [30][31] show how SVM classifier for Ordinal data is better when compared to Ordinal Logistic and Multinominal Logistic Regression. Also, the results from this paper show an improvement of SVM accuracy when compared with Ordinal logistic/proportional odds model [32] when applied to mutual fund data with imbalance

In this paper, we have organized the discussion as below: A overview of the literature on class imbalance, SVM, and ordinal data is given in Section 2. The solution approach is in Section 3, which also discusses the mutual fund dataset and the framework for the suggested model. Section 4 presents and discusses the study's findings. The work is finally concluded in Section 5, which also discusses its shortcomings and makes suggestions for more research



1. PROPOSED METHODOLOGY

In this research work ,we attempt to evaluate SVM as a classifier for imbalanced ordinal class data and test the same on how the classifier can accurately rate the Indian mutual fund. Here we run an SVM classifier -a Supervised machine learning algorithm on selected 98 large cap mutual funds from Indian market and compare and validates if the given ratings and the ML generated ratings fall in the same class. The model is also tested for two real world data sets with before and after resampling techniques is applied to the datasets. The steps followed are as follows.

A. Data collection and Resampling: - For this study we have taken a data set of 98 large

cap equity mutual funds from Value research online website. The data set has information on financial ratios of each fund, such as Alpha, Beta, Sortino Ratio, Standard Deviation, Sharpe Ratio, and R-Squared along with ratings given by the analyst. Further among the 98 funds the rating distribution is as follows: 7 funds have Rating 5, 24 funds have Rating 4, 31 funds have Rating 3, 26 funds have Rating 2, and 10 funds have Rating 1.There is clear evidence of class imbalance in this dataset with class 4,3, and 2 being majority classes and class 5 and 1 being minority classes thus any machine learning model would tend to favor majority class proportion of observations reducing the accuracy. Therefore for any imbalanced datasets, a correct balance of class distribution is required for better classification of ratings.

In general, there are some approaches to handling class imbalance problems which has been already discussed in paper(30) and we apply the same resampling methods using and hybrid strategy or a mixed class balance strategy method EM+GK Means where the EM algorithm(Dempster et al., 1977) is used for under-sample the majority classes and GK means (M. M. Hassan et al, 2021) is used to oversample or generate more instance of the minority class. We then compare between results of the SVM classifier with Linear and Non-Linear kernels on imbalanced data, and balanced data using EM+GK Means and conclude the accuracy of the classifier. R software is used to implement the methods.

B. Model Building and Results

In this paper, Support vector machine algorithm is used to classify the funds into its rating category and notice that the accuracy is very poor for the imbalanced data. Hence to reduce the data skewness, we use the EM algorithm [3] to undersample majority classes 4,3 and 2 and oversample minority classes 5 and 1 using SMOTE accordingly. We then run the SVM model.

Confusion Matrix for Imbalanced Class

	1	2	3	4	5
1	5	0	3	0	0
2	0	3	2	0	0
3	0	0	9	29	0
4	0	0	1	37	0
5	0	0	0	8	35

Table- I: Evaluation Criteria for imbalanced class

Evaluation Metrics \ Class	1	2	3	4	5
Sensitivity	1.00	1.00	0.6	0.5	1.00
Specificity	0.97	0.98	0.75	0.98	0.91
Precision	0.62	0.60	0.23	0.97	0.81
Recall	1.00	1.00	0.93	0.60	1.00
Balanced Accuracy	0.98	0.79	0.67	0.74	0.95

EM and GK means: Due to its tendency to cause over-generalization when faced with a substantial class imbalance and its limitations with high-dimensional data, the SMOTE

algorithm in SCUT necessitates an alternative approach. Addressing these concerns, M.M. Hassan proposed hybrid GK-Means- an method to oversample based on Gaussian distribution and K-means method in 2001, demonstrating superior performance compared to SMOTE.

Considering this, we introduce a novel hybrid class balancing technique that merges GK-Means for oversampling with the EM algorithm for under sampling. This combined approach aims to mitigate the class imbalance problem in our dataset more effectively. Since we have 98 samples in the original data set, we take a mean of 20 samples for each class. For minority classes of 1 and 5, we use GK to oversample the data and create 20 instances of each class. For majority classes 2,3 and 4 we use EM algorithm to draw 20 samples from each class and then merge all the classes. Now we have 100 instances with 20 samples from each class thus having a balanced dataset. Now on using K means clustering model on the new data set, we observe that the new class balancing technique of combining EM and GK means gives a better accuracy of 84% when compared to SCUT technique that gives only 77% accuracy while the clustering accuracy on imbalanced classes is 67% accuracy. Also, from below Table-III we notice that the various classification measures such as Specificity, Sensitivity, Recall Precision, and F1 measures of the EM+GK means show higher performance than SCUT. These results conclude that the proposed class balancing method solves the issue of class imbalance and K Means algorithm can now cluster datapoints more accurately.

Table- III: Evaluation Metrics for SMOTE and EM

Evaluation Metrics\Class	**1**	**2**	**3**	**4**	**5**
Sensitivity	1.00	0.9	0.88	0.74	1.00
Specificity	1.00	0.97	0.93	1.00	0.95
Precision	0.612	0.9	0.75	1.00	0.80
Recall	1.00	0.97	0.97	0.91	1.00
Balanced Accuracy	1.00	0.93	0.91	0.87	0.97

Table IV below displays the overall accuracy p-value and class intervals for each class balancing technique. The significance of the p-values is determined based on a threshold of $p<0.05$, indicating statistical significance. Therefore, we can conclude that our results are statistically significant and valid.

Table-IV: Metrics table for accuracy, p-values, Confidence intervals for each Class Balancing Technique

	Class Balancing Techniques			
Evaluation Metric\Kernel	**Imbalanced Class**		**EM and GK Means**	
	SVM (Linear Kernel)	SVM(RBF Kernel)	SVM (Linear Kernel)	SVM(RBF Kernel)
Accuracy (%)	65	59	85	90
p-value	0.04337	0.1098	0.0000007	0.00000032
95% CI	(0.3833,	(0.3292,	(0.6211,	(0.683,

	0.8579)	0.8156)	0.9679)	0.9877)
No Information Rate	0.4118	0.4118	0.2	0.2

We then apply the SVM classifier with hybrid resampling technique on Wine and Glass data sets. These datasets have an ordinal output, and we notice the SVM classifier with our hybrid resampling technique have a better accuracy compared to imbalanced dataset. Also, the RBF kernel shows an improvement on accuracy over Linear. These results confirm that when SVM classifier is used with EM and GK Gauss as resampling techniques, the results show an improvement.

		Imbalanced class Accuracy(in %)		**Balanced class Accuracy(in %)**	
Datasets	# of classes	SVM(Linear Kernel)	SVM(RBF Kernel)	SVM (Linear Kernel)	SVM(RBF Kernel)
wine	6	57	67	59	69
Mutual funds	5	65	59	85	90
Glass	6	63	70	66	74

2. FINAL REMARKS AND FUTURE WORK

In conclusion, this study successfully compares the hybrid Support Vector Machine (SVM) model with ordinal logistic regression in handling imbalanced data, demonstrating the strengths and limitations of each approach. The hybrid SVM model outperforms ordinal logistic regression in certain imbalanced data scenarios, offering a more robust solution for predicting ordinal outcomes. Furthermore, the study's extension to predicting mutual fund ratings showcases the practical applicability of these methods in real-world financial data.

For future work, further refinement of the hybrid SVM approach could be explored, particularly in handling even more complex or large-scale datasets. Additionally, extending the model to other industries and datasets could provide deeper insights into its versatility and robustness. Further comparisons with other machine learning techniques could also help in identifying the most optimal model for various imbalanced classification problems.

REFERENCES

1. V. Vapnik, The nature of statistical learning theory. Springer-Verlag New York, Inc., 1995.
2. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
3. B. Boser, I. Guyon, and V. Vapnik, “A training algorithm for optimal margin classifiers,” in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144–152, ACM Press, 1992.
4. N. Cristianinio and J. Shawe-Taylor, An introduction to support Vector Machines: and other kernel-based learning methods. Cambridge University Press, 2000.
5. B. Scholkopf and A. Smola, Learning with Kernels: Support Vector Machines,

Regularization, Optimization, and Beyond. Cambridge, MA, USA: MIT Press, 2001.
6. C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121–167, 1998
7. A.Astha, V.L.Herna & P. Eric, (2015), "SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling", KDIR, SciTePress
8. Ben-David, I., J. Li, A. Rossi, and Y. Song. 2021. What Do Mutual Fund Investors Really Care About? Review of Financial Studies, forthcoming.
9. Chawla,N.,Bowyer, K,Hall and Kegelmeyer,W.P. (2002)."Smote:synthetic minority over sampling", Journal of artificial intelligence research, vol.16(1)
10. Cheng, Si and Lu, Ruichang and Zhang, Xiaojun, What Should Investors Care About? Mutual Fund Ratings by Analysts vs. Machine Learning Technique (2021).
11. Choi, J. J., and A. Z. Robertson. 2020. What Matters to Individual Investors? Evidence from the Horse's Mouth. Journal of Finance 75:1965−2020
12. D. Acharya, G. Sidana, "Classifying mutual funds in India: Some results from Clustering", Indian Journal of Economics and Business, vol. 6, no. 1, pp. 71-79, 2007.
13. Hereil, Pierre and Mitaine, Philippe and Moussavi, Nicolas and Roncalli, Thierry, Mutual Fund Ratings and Performance Persistence (June 25, 2010).
14. Hsu C.F, Hung H.F. 2009. Classification methods of Credit Rating – A Comparative Analysis on SVM, MDA and RST. Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on: 1-4. DOI: 10.1109/CISE.2009.5366068
15. Jaime, Cardoso and Joaquim. d. Costa, (2007), "Learning to classify ordinal data: the data replication method," Journal of Machine Learning Research, vol. 8
16. Lee, Hansoo & Kim, Jonggeun & Kim, S. (2017). "Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions". International Journal of Fuzzy Logic and Intelligent Systems. 17. 229-234.
17. Lisi, Francesco & Otranto, Edoardo. (2008). "Clustering Mutual Funds by Return and Risk Levels". Mathematical and Statistical Methods for Actuarial Sciences and Finance. Vol 10. 978-988
18. M. M. Hassan, A. S. Eesa, A. J. Mohammed and W. K. Arabo, "Oversampling method based on gaussian distribution and k-means clustering," Computers, Materials & Continua, vol. 69, no.1, pp. 451–469, 2021.
19. Marathe, Achla & Shawky, Hany. (1999). "Categorizing mutual funds using clusters". Advances in Quantitative Analysis of Finance and Accounting. 7.
20. McCullagh(1980),"Regression models for ordinal data," Journal of the Royal Statistical Society, Series B, vol. 42
21. Michael C. Jensen, (1968), "The Performance of Mutual Funds in the period 1945-1964', The Journal of Finance
22. R. Evans, B. Pfahringer and G. Holmes, "Clustering for classification," 2011 7th International Conference on Information Technology in Asia, 2011
23. Raina, Battle, Packer, and A. Ng. Self-taught learning: transfer learning from unlabeled data. In ICML, 2007.
24. S. Takumasa, M. Tohgoroh, Mutoh, Atsuko, Inuzuka, Nobuhiro. (2015). "Clustering Mutual Funds Based on Investment Similarity". Procedia Computer Science.60.881-

890

25. R. Herbrich, T. Graepel, and K. Obermayer. Support vector learning for ordinal regression. In Proc. of 9th International Conference on Artificial Neural Networks (ICANN), pages 97–102. 1999.
26. W. Chu and S.S. Keerthi. New approaches to support vector ordinal regression. In Proc. of International Conference on Machine Learning (ICML-05), pages 145–152. 2005. [12]
27. W. Chu and S.S. Keerthi. Support vector ordinal regression. Neural Computation, 19(3), 2007
28. Aroef, Chelvian & Yuda, Rivan & Rustam, Zuherman & Pandelaki, Jacub. (2019). Multinomial Logistic Regression and Support Vector Machine for Osteoarthritis Classification. Journal of Physics: Conference Series. 1417. 012012. 10.1088/1742-6596/1417/1/012012.
29. Eviyana Atmanegara1 , Taly Purwa (2021). Hybrid Support Vector Machine and Logistic Regression for Multiclass Classification: A Case Study on Wine Dataset. Indonesian Journal of Data Science Vol.1, No.1, April 2021, pp. 1~7
30. Ashoka and Ismail B,(2020), Mutual Fund Rating Prediction using Proportional Odds Logistic Regression with Imbalanced Class, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5S, January 2020
31. S. Rajaram, A. Garg, X. S. Zhou, and T. S. Huang, "Classification approach towards ranking and sorting problems," in Proc. of the 14th European Conference on Machine Learning (ECML), ser. Lecture Notes in Computer Science, vol. 2837, 2003, pp. 301–312.
32. Caruana, S. Baluja, and T. Mitchell. Using the future to sort out the present: Rankprop and multitask learning for medical risk evaluation. In Advances in neural information processing systems 8 (NIPS). 1996.
33. P. McCullagh and J. A. Nelder. Generalized Linear Models. Chapman and Hall, London, 1983.
34. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:237–297. [Google Scholar]
35. Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. Boston: MIT Press; 2002.