High Dimensional Unbalanced Data Classification Vs SVM Feature Selection

S. Chinna Gopi¹, B. Suvarna² and T. Maruthi Padmaja²*

¹Vijaya Institute of Technology for Women (VITW), Enikepadu, Vijayawada - 521108, Andhra Pradesh, India; Simhadri594@gmail.com
²VFSTR University, Vadlamudi, Guntur - 522213, Andhra Pradesh, India; sv1720@gmail.com, Padmaja.tu2002@gmail.com

Abstract

Background/Objectives: It is well known that the performance of the classification models prone to the class imbalance problem. The class imbalance problem occurs when one class of data severely outnumbered the other classes of data. The classification models learned on Support Vector Machines (SVM) are quite prominent in exhibiting better generalization abilities even in the context of the class imbalance problem. However, it is proved that the high imbalance ratio hinders SVM learning performance. With this concern, this paper presents an empirical study on the viability of SVM in the context of feature selection from moderately and highly unbalanced datasets.

Methods/Statistical Analysis: The Support Vector Machine-Recursive Feature Elimination (SVM-RFE) wrapper feature selection is analyzed in this study and its performance on one document analysis and two biomedical unbalanced datasets is compared with two prominent feature selection methods like Chi-Square (CHI) test and Information Gain (IG) using Decision Tree and Naïve Bayes classification models.

Findings: From this empirical study two major identifications are reported: 1. For the considered scenarios, classification models learned on IG and CHI test are better performed than SVM-RFE feature selection of high class imbalance setting. 2. The SVM-RFE on rebalanced data yielded better performance than SVM-RFE on original data.

Application/Improvements: Considered feature selection methods, including SVM-RFE yielded better performance on oversampled data than SVM-RFE on original data. Overall, this study reports models learned on Decision Tree exhibited better performance than the models learned on Naïve Bayes classifier.

Keywords: Class Imbalance Problem, Chi-Square, Information Gain, Support Vector Machine, SVM-RFE

1. Introduction

The performance of the classification models on real world problems like Crime Detection, Intrusion Detection, Rare Disease prediction is hindered by the class imbalance problem¹-⁴. The class imbalance problem occurs when one class of data severely under represented than the other classes of data such that the underrepresented minority class cannot be predicted by the classification models. Usually in the context of class imbalance, the performance of the classification model which is learned with any of the algorithms, like Decision Tree (DT), k-Nearest Neighbor (kNN), Neural Networks (NN), Naïve Bayes (NB) biased towards the majority class. From the experimental studies ⁵, the Support Vector Machines are less sensitive to the class imbalance problem when compared with other classification algorithms. However⁶ proved that the heavy class imbalance renders SVM boundary skewed towards minority class samples. To ameliorate the class imbalance problem, several solutions are proposed at the data level and at classification algorithm level. Ensemble techniques like bagging, boosting and meta learning is also adopted to improve the classifier performance in the case of class imbalance.

According to⁷-⁸, the hardness of the class imbalance problem is due to data characteristics like dataset size, small disjuncts, and class overlap of a dataset. As per⁹, the number of features also play key role in performance

*Author for correspondence
degradation. Consequently, adapting feature selection is a beneficiary to model performance while handling high dimensional unbalanced datasets\textsuperscript{10}. Therefore, feature selection becomes crucial for high dimensional class imbalance problems.

Support Vector Machine-Recursive Feature Elimination (SVM-RFE) is one of the prominent feature selection methods and widely applied to biomedical datasets\textsuperscript{11}. The SVM-RFE algorithm works based on the criterion of removal of a feature minimizes the variation in $\|w^T\mathbf{F}\|$, where \( w \) is the width of the hyperplane. From the studies the viability of adopting SVM-RFE feature selection of highdimensional unbalanced datasets is a debatable issue. To address this issue, in this work, like an empirical study is carried out on unbalanced and balanced datasets over 3 real world highdimensional datasets. The balancing of unbalanced datasets is performed by adopting resampling techniques like Random Undersampling and Random oversampling on original data. Further, the performance of a SVM_RFE feature selection is compared with feature selection measures like Information Gain, Chi_Square over balanced and unbalanced data. The whole empirical study is carried out over classification models learned on DT and NB algorithms. The performance measure Area under ROC (AUC) is used to evaluate the classifiers learned on different feature selection methods. This study presents two major identifications regarding to effect of class imbalance on SVM feature selection:

- On highly highdimensional unbalanced datasets on an average SVM-RFE under performs than the other feature selection methods due to the skewed boundary towards the minority class.
- SVM-RFE on the balanced class distributions yielded better performance than the unbalanced distributions.

Many solutions were proposed at data level, classification algorithm and the hybrids of both, to combat with this class imbalance problem. All these mentioned solutions adopt sampling strategies, prediction weighing, ensembles and novelty detection methods either to balance or to weigh the data space. But actually the down fall of the classification model performance in the context of unbalanced datasets is not exact due to class under representation, but due to the data characteristics like dataset size, concept complexity, degree of overlapping, number of features and classification algorithm being used aggravates the minority class prediction. According to number of features in the highdimensional unbalanced data is also one of the key reasons for classification model performance degradation and feature selection is one of the key aids to combat with this problem. In this context several attempts were made to identify the viability of feature selection methods and to identify best feature selection methods for high dimensional class imbalance problems\textsuperscript{12-14} like text classification and microarray data classification. The major finding from these studies is that adopting feature selection methods improved the performance of the unbalanced data classification problems. But none of the findings matches with others and these findings mostly depend on the classification algorithm, feature selection methods and the classification performance measures that are considered for the particular study. As a follow-up on this line, the current work mainly focuses to carry out an empirical study in order to identify the viability of SVM-RFE over unbalanced datasets.

Further, new feature selection methods are devised specifically to combat with the highdimensional class imbalance problem, suggested to use combined features minority and majority class with one-sided and two sided feature selection methods. Recently, a feature selection algorithm proposed\textsuperscript{15} named Feature Assessment by Sliding Thresholds (FAST) for small sample and unbalanced data classification problems. This algorithm evaluates a linear classifier over single feature along multiple thresholds and the features with the greatest area under ROC are selected. An another method named Feature Selection for Minority Class (FSMC)\textsuperscript{16} selects those features whose minority class mean value is a minimum of two standard deviations away from the majority class.

2. Background

This section depicts the background needed to carry out the empirical study to identify the effect of class imbalance on Support Vector Recursive Feature Selection (SVM_REF).

2.1 Support Vector Machine

The Support Vector Machine (SVM)\textsuperscript{17} is a binary classification algorithm and exhibited better generalization abilities on many real world problems. The SVM classifier describes an optimal hyperplane in feature space, given a training set of \( N \) data points \( \{(x_i, y_i)\}_{i=1}^{N} \), with
input data \( x_i \in \mathbb{R}^n \) and corresponding binary class labels \( y_i \in \{+1, -1\} \).

Figure 1. Support vector machine classification.

The optimal hyperplane that separates the two classes by the largest margin drawn on a global optimization criteria that minimizes

\[
\min_{w, b, \xi} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i
\]

subject to

\[
\begin{align*}
    y_i (w^T x_i + b) \geq 1 - \xi_i \\
    \xi_i \geq 0
\end{align*}
\]

Where \( w \) is the norm of the hyperplane, \( b \) is the intercept of hyperplane from the origin, \( x_i \) is the input vector to feature space, \( y_i \) is the corresponding class label, \( \xi \) slack variable for handling nonlinearity and \( C \) is the tuning parameter for corresponding loss function for misclassification cost. Generally, Equation (1) and (2) are solved by convex Quadratic Programming (QP) Problem by formulating its dual cost function given as

\[
\max w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j)
\]

subject to

\[
\begin{align*}
    y_i \alpha_i \leq C \\
    \sum_{i=1}^{N} \alpha_i y_i = 0
\end{align*}
\]

Here \( \alpha_i \)'s are the Lagrange multipliers whose values are non-zeros for the training instances that are at the margin. Those training instances are known as support vectors. In Equation (3), \( k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \), represents the positive definite kernel that satisfies Mercers theorem. Here \( \Phi(x_i) \) is the mapping of feature space for \( x_i \). After solving QP the norm of the hyperplane can be represented as

\[
w = \sum_{i=1}^{N} \alpha_i y_i \Phi(x_i) \]

and the test instances are classified with

\[
y(x) = \text{sign} \left[ \sum_{i=1}^{N} \alpha_i y_i \Phi(x, \Phi(x_i)) + b \right]
\]

Equations (5) and (6) show the significant role of support vectors in defining SVM boundary. Figure 1 depicts the SVM optimal boundary of separation between the two classes. As per the optimal boundary of the SVM model learned on highly imbalance datasets is skewed towards the minority class leaving elongated boundary towards the majority class. Therefore the SVM model learns on unbalanced data biased towards the majority class prediction. Figure 2 depicts the skewed boundary towards minority class at high class imbalance scenarios.

Figure 2. SVM classifier with skewed boundary on highly unbalanced data.

2.2 Feature Selection

Feature selection methods to select those features that are relevant to the classification algorithm from the given dataset and the feature selection methods that are considered for the study are described below.

- **Support Vector Recursive Feature Elimination Algorithm (SVM-RFE):** The main objective of SVM_RFE algorithm is to find out \( r \) features, among \( n \) given features where \( (d < n) \) which maximizes the classification model performance. This is a backward sequential feature selection
algorithm based on the criterion of removal of a feature minimizes the variation in $\|\Phi(x)\|^2$. Initially, the SVM_RFE algorithm starts with all $n$ features of the dataset and incrementally removes one feature after another until the feature list is empty. Each feature is ranked based on the evaluation criterion given in Equation (7). The best feature $k$ is selected for removal is based on $\min R_i(k)$, this $R_i(k)$, is calculated as

$$R_i(k) = \|\Phi(x_i)\|^2 = \sum_{i=1}^{N} \alpha_i^2 \Phi(x_i)$$  \hspace{1cm} (7)$$

Pseudo-Code 1 depicts the algorithm for SVM-RFE.

Pseudo-Code 1. SVM-RFE algorithm

Input: $Var$ = {1, 2, 3, … n}

Rank = {};

Repeat

Train an SVM classifier using all variables $Var$ = {1, 2, 3… n} on the training data.

Rank the attribute $k$ based on the evaluation criteria $R_i(k)$, for all the variables of $Var$.

$best\_rank = \min R_i(\cdot)$;

Rank = Rank $\cup Var[best\_rank]$;

$Var$ = $Var[1,2,3,....best\_rank-1,best\_rank+1,....n}$

Until $Var$ is empty.

2.3 Resampling Methods

As the resampling methods are quite widely used for solving real world class imbalance problems in this study, we have adopted random under sampling and oversampling to balance the unbalanced datasets. Random oversampling balances the class distribution by replicating random minority class samples. Random undersampling balances the class distribution by eliminating random majority class samples. In the literature, there are many oversampling and undersampling techniques to improve the classifier performance in the context of class imbalance. As these methods represent the same notion of oversampling and undersampling we have limited our study to random undersampling and random oversampling.

3. Experimental Results and Discussion

This section presents the details of datasets considered and the discussion on the results obtained in a detailed experimental study conducted to analyze the behavior of SVM-RFE over DT and NB classifier. Two other feature selection methods IG and CHI are used to compare the performance of SVM-RFE over unbalanced data.

**Datasets**: The datasets from two different domains, two biological datasets and one text recognition dataset is used to study the effect of class imbalance on SVM-RFE feature selection algorithm.

The statistics for the considered datasets is presented in Table 1 and the detailed description of each dataset is given below:

- **Information Gain (IG)**: The difference between entropies of the class distributions is measured by IG and the entropy measures the impurity of the given variable. The mathematical notation for the IG is given below

$$IG = \text{Entropy}(\text{Pos, Neg}) - \left(\frac{TP \times PD}{N} \times \text{Entropy}(\text{TP, PD}) + \frac{FN \times FN}{N} \times \text{Entropy}(\text{TN, FN})\right)$$  \hspace{1cm} (8)$$

Here TP is the True positive rate that the positive class classified correctly. FP is the false positive rate of negative class misclassification. TN is the true negative rate of negative class samples classified correctly; FN is the false negative rate of positive class misclassification.
Table 1. Datasets description

| Dataset                  | No. of Features | No. of Samples | Imbalance Ratio | POS, NEG        |
|--------------------------|-----------------|----------------|-----------------|----------------|
| Central Nerves System_c  | 7129            | 60             | 2.35            | Survivors, Failures |
| TIS-5%                   | 927             | 668            | 3.6             | Positive, Negative |
| page-block0              | 10              | 5472           | 8.79            | Rest, Text      |

Central Nerves System_c (CNS_c): This dataset describes the predictive problem of Survivors from Central nervous system Embryonal Tumor. The dataset contains two classes’ survivors, failures from Embryonal tumo.

Translation Initiation Sites (TIS): This is the imbalanced version of the original TIS dataset and describes to distinguish true TIS with false TIS samples. As this dataset is huge enough of having 13375 samples, we have synthesized this to 668 samples as described in.

3.1 Experimental Setup
In order to conduct experiments SVM-RFE algorithm is implemented using LIBSVM’s MATLAB MEX interface and the rest of the methods IG, CHI, NB, DT are adopted from WEKA. As the Classification accuracy is not adaptable to evaluate the classifier performance, an unbiased measure Area Under ROC is used as performance measure to evaluate the classifier performance over DT and NB.

3.2 Results and Discussion
The DT and NB classifier performance drawn on after applying SVM-RFE, IG and CHI over CNS_c, TIS, page-block0 datasets are depicted in Figures 3-5. In the figures the entry starts with ov represents the results from oversampling and un represents the results from undersampling and the rest are drawn on original unbalanced datasets.

From the obtained results over three datasets it can be observed that all together SVM-RFE feature selection on DT and NB classifiers is underperformed than IG and CHI feature selection methods. For CNS_c (Figure 3,) dataset the classifier performance in terms of AUC for top 100 ranked features over DT and NB classifiers SVM_RFE yielded lower AUC when compared with IG and CHI. Similar kind of results is observed for TIS dataset regarding to SVM-RFE performance (Figure 4). But for page-block0 dataset different results from earlier datasets are observed (Figure 5). For this dataset the result for top 5 features over a DT classifier on all considered feature selection methods yielded similar performance, whereas over NB classifier up to 4 features, SVM-RFE exhibited lower AUC, at the 5th feature, it has yielded a similar AUC with IG and CHI. Thus, from these results, it is clearly discernible that SVM-RFE underperformed on highly unbalanced datasets like CNS+c and TIS and yielded better performance on a moderate unbalanced dataset page-block0 compared with IG and CHI. This finding regarding to SVM-RFE mimicking the conclusions regarding to the findings of the effect of class imbalance on SVM classifier. As the objective of SVM-RFE is to find out the best variables that reduces the variation in hyperplane regarding to this $\| \cdot \|^2$, is skewed towards the minority class (Figure 3,4), thus may leads to poor AUC compared to IG and CHI.

Figure 3. Average ROC performance on CNS_c data over decision tree and naïve bayes classifiers.

Figure 4. Average ROC performance on TIS data over decision tree classifier naïve bayes classifiers.

Figure 5. AUC performance on page-block0 data over decision tree and naïve bayes classifiers.
But interestingly SVM-RFE on oversampled data over DT and NB yielded better AUC compared with SVM-RFE on original data alone. Further, classifiers learned on IG and CHI on oversampled data exhibited better performances compared to the AUCs on original data. In this scenario, for the three datasets the best AUC reported by SVM-RFE on the classifiers is either better or competitive with the AUC performances of IG and CHI. Thus, this evidence gives the insight that balancing the original data could improve the SVM-RFE feature selection on the class imbalance problem.

However, for the considered datasets and for all feature selection methods, random underampling of the data is not able to improve the AUC performance on both the classifiers. Further, the AUC graphs for the random under sampling are in lag behind the performance of the considered feature selection methods on original data. This might be due to smaller sizes of data considered for the experimental study and undersampling the majority class further leads to more skewed distributions.

For all the datasets, it is identified that the feature selection methods IG and CHI performed equally well on balanced and unbalanced data. No clear winner is identified among them from the considered study. From all the experiments with three feature selection methods on three datasets over two classifiers the best performance is reported by SVM-RFE on oversampled data. In this case, though for many of the results the AUC on SVM-RFE at initial best features seems to be lags behind the Ovig and OVCHI as the number feature increases the AUC got improved and yielded a better AUC than IG and CHI square methods. Additionally, in this study it is noticed that DT has exhibited clear dominance in AUC than NB for the three datasets over the balanced and unbalanced data.

4. Conclusion

Feature selection methods improve the classifier performance while handling high-dimensional unbalanced datasets. This study explored the viability SVM-RFE feature selection algorithm in the context of high-dimensional, highly unbalanced datasets. SVM-RFE designed based on the principles of discarding the variables that are minimizing the variation in current hyperplane, which is skewed to minority samples in the case of highly unbalanced datasets. This study leads to the following conclusions regarding the adoption of SVM-REF feature selection:

- For highly unbalanced, highdimensional datasets SVM-REE underperformed when compared to IG, CHI feature selection methods.
- Adopting SVM-RFE on rebalanced distribution improved the classifier performance.
- SVM-RFE on random oversampling yielded better performance than random undersampling.

5. References

1. Iqbal R, Murad MAA, Mustapha A, Panahy PHS, Khanamadiliravi N. An experimental study of classification algorithms for crime prediction. Indian Journal of Science and Technology. 2013 Mar; 6(3):4219–25.
2. Sivaranjani S, Sivakumari S. A novel approach for serial crime detection with the consideration of class imbalance problem. Indian Journal of Science and Technology. 2015 Dec; 8(34):1–9.
3. Sun Y, Wong AC, Kamel MS. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence. 2009; 23(4):687–719.
4. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009 Sep; 21(9):1263–84.
5. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intelligent Data Analysis Journal. 2002 Oct; 6(5):429–9.
6. Wu G, Chang E. Class-boundary alignment for imbalanced dataset learning. Proceedings of ICML’2003 Workshop on Learning from Imbalanced Data Sets II; Washington DC. 2003 Aug. p. 1–8.
7. Jo T, Japkowicz N. Class imbalances versus small disjuncts. ACM SIGKDD Explorations. 2004 Jun; 6(1):40–9.
8. Prati RC, Batista GEAPA, Monard MC. Class imbalances versus class overlapping: An analysis of a learning system behavior. Proceedings of Mexican International Conference on Artificial Intelligence; 2004 Apr. p. 312–21.
9. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks. 2008 Mar-Apr; 21(2–3):427–36.
10. Wasikowski M, Chen XW. Combating the small sample class imbalance problem using feature selection. IEEE Transactions on Knowledge and Data Engineering. 2010 Oct; 22(10):1388–400.
11. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning Journal. 2002; 46:389–22.
12. Mladenic D, Grobelnik M. Feature selection for unbalanced class distribution and naive bayes. Proceedings of 16th Int’l Conference on Machine Learning; 1999. p. 258–67.
13. Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter. 2004 Jun; 6(1):80–9.
14. Bolon-Canedo V, Sanchez-Maroña A, Betanzos A, Benítez JM, Herrera F. A review of microarray datasets and applied feature selection methods. Information Science. 2014 Oct; 282:111–35.
15. Chen X, Wasikowski M. FAST. A ROC-based feature selection metric for small samples and imbalanced data classification problems. Proceedings of ACM SIGKDD; 2008. P. 124–33.
16. German C, Angelica M.M, Eduardo FM. A minority class feature selection method. Proceedings of Pattern Recognition, Image Analysis, Computer Vision, and Applications. 2011 Nov; 7042:417–24.
17. Vapnik V. Statistical learning theory. Wiley–Interscience; 1998.
18. Chawla N, Bowyer K, Hall L, Kegelmeyer P. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. 2002; 16:321–57.
19. Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Computational Intelligence. 2004 Feb; 20(1):18–36.
20. Padmaja TM, Narendra D, Bapi RS, Krishna PR. Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection. Proceedings of 15th International Conference on Advanced Computing and Communications; 2007 Dec. p. 511–6.
21. Thongkam J, Xu G, Zhang Y, Huang F. Toward breast cancer survivability prediction models through improving training space. Expert Systems with Applications. 2009 Dec; 36(10):12200–9.
22. Alcala-Fdez J, Fernandez A, Luengo J, Derrac J, Garcia S, Sanchez L, Herrera F. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing. 2011; 17(2–3):255–87.
23. Nervous System. Available from: https://en.wikipedia.org/wiki/Nervous_system
24. Levis Tonji Education-Mirror Kentridge. Available from: http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html#TIS
25. Santiago DV, Cunningham P. An evaluation of dimension reduction techniques for one-class classification. Artif Intell Rev. 2007 Apr; 27(4):273–94.
26. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2(27):1–27.
27. Mark H, Eibe F, Geoffrey H, Bernhard P, Peter R, Ian HW. The WEKA data mining software: An update. SIGKDD explorations. 2009 Jun; 11(1):10–8.
28. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters-Special Issue: ROC Analysis in Pattern Recognition. 2006 Jun; 27(8):861–74.