DOI: 10.21917/ijsc.2021.0358

FEATURE SUB-SPACING BASED STACKING FOR EFFECTIVE IMBALANCE HANDLING IN SENSITIVE DATA

S. Josephine Theresa and D.J. Evanjaline
Department of Computer Science, Bharathidasan University, India

Abstract

Several real world classification applications suffer from an issue called data imbalance. Handling data imbalance is crucial in developing an effective classification system. This work presents an effective classifier ensemble model, Feature Sub-spacing Stacking Model (FSSM) that has been designed to operate on highly imbalanced, complex and sensitive data. The FSSM technique is based on creating subspace of features, to aid in the reduction of data complexity and also to handle data imbalance. First level trains models based on these features, which is followed by creating a stacking architecture. The second level stacking architecture trains on the predictions from the first level base models. This has enabled better and qualitative predictions. Experiments were conducted on bank data and also the NSL-KDD data. Results reveal highly effective performances compared to the existing models.

Keywords:

Classification, Data Imbalance, Ensemble, Stacking, Feature Sub-Spacing

1. INTRODUCTION

Classification plays a vital role in data analysis and predictions in several fields. Some of the major applications include detection of credit or debit card frauds, intrusion detection in networks, medical diagnostics, etc. Imbalance is a major issue in all these fields. Data is imbalanced if one class contains large number of instances compared to other existing classes in the data [1]. The class that exhibits large number of instances is considered to be the majority class, while all the other classes are considered to be minority classes. Training classification models on imbalanced data becomes complicated due to the bias introduced by the majority classes. Correct prediction of minority classes is affected due to imbalance. In real world applications, prediction of minority classes is more crucial compared to the majority classes [2]. For example, in detecting credit card frauds, fraudulent data represents the minority class, however, detecting frauds is more important compared to detecting legitimate transactions. Similarly in network intrusion detection, detecting the intrusion signatures is much more important compared to detecting normal signatures. This shows the significance of handling data imbalance during the classification process.

Several techniques have been proposed in literature to handle data imbalance. The proposed techniques can be broadly classified into data handling methods [3] [4]; algorithm based methods [5] and combination methods that uses both these techniques [6]. Data handling methods are mainly centered on sampling techniques [7]. Oversampling and under sampling are the two major sampling techniques. Oversampling deals with creating minority instances from existing minority data, while under sampling deals with pruning the majority instances to ensure balance in data. Recent works like SMOTE [8] are based on applying both the sampling methods to obtain balanced data. Algorithm based methods include cost or weight based analysis [9] [10] that increases the weight of minority class instances to enhance the training process. Ensemble based methods also fall under this category.

This work proposes an ensemble based classifier model to handle the data imbalance and the data complexity. Multiple bags are created to reduce the imbalance levels. Stacking based methodology is incorporated to handle data with high complexity levels during prediction.

Remaining of this work is structured as follows; section 2 presents the related works, section 3 presents a detailed view of the proposed FSSM technique, section 4 presents the results, section 5 presents the comparisons, and section 6 concludes the work.

2. RELATED WORKS

Handling data imbalance during classification is a major requirement in most systems operating on data generated in real-time. Several techniques have been researched and proposed to handle the issue of data imbalance. This section discusses some of the contributions proposed to handle data imbalance.

Data rebalancing is one of the major research areas concentrating towards handling data imbalance. A Bayesian Neural learners' based model focussing on rebalancing to handle data imbalance was proposed by Lazaro et al. [11]. The technique uses binary Bayesian classifiers, which exhibit intrinsic re-balancing properties. The intrinsic rebalancing is made possible by using the diversification mechanism which is based on applying varying cost policies to the model.

An ensemble model based on data balancing and dynamic selection of machine learning models is proposed by Gao et al. [12]. Balancing of data is performed by data partition hybrid sampling technique. Data is partitioned based on the level of imbalance, and oversampling is applied for balancing. Three ensemble models are created, and a dynamic selection rule is designed to select the appropriate model based on the data.

A random under sampling model for data balancing was proposed by Sui et al. [13]. The model performs data balancing by removing instances from the majority class in-random. This, however leads to loss of information. Other similar resampling techniques include one-sided selection model by Zuo et al. [14] and an under sampling approach in [15].

An incremental model for handling data imbalance in credit card transactions was proposed by Somasundaram et al. [16]. This technique creates data divisions by considering temporal aspects of the data to produce effective results. A neural network based model that is used to handle data imbalance was proposed by Suh et al. [17]. A classification enhancement generative adversarial

networks was proposed in this work. The model is used to generate minority classes to improve the prediction performance. A Weighted Extreme Learning Machine (WELM) based model for imbalance classification was proposed by Zhu et al. [18]. This work mainly aims towards achieving effective performance by applying multiple optimization methods to optimize the WELM model to achieve the desired performance. Optimizing the weights of neural networks to achieve desired performances has been under major research in imbalance data handling. Some such optimization models include works by Zhu et al. [19], Cao et al. [20], Xu et al. [21] and Han et al. [22].

An ensemble based model that uses weight based techniques for improving classification accuracy on imbalanced data was proposed by Wang et al. [23]. This work combines the XGBoost ensemble with weighted and focal losses. The inclusion of these modules are stated as the major reasons that can effectively tackle imbalance in classification data. Other similar techniques that are used to tackle imbalance in ensemble based models include works by He et al. [24], Liu et al. [25] and Kabir et al. [26].

3. FEATURE SUB-SPACING BASED STACKING MODEL (FSSM)

Sensitive data tends to be highly complex in nature. Highly complex data requires highly complex architectures. Data imbalance further creates bias in the training models. In order to reduce the impact of data imbalance and to handle the complexity contained in the data, this work proposes a Feature Sub-spacing based Stacking Model (FSSM).

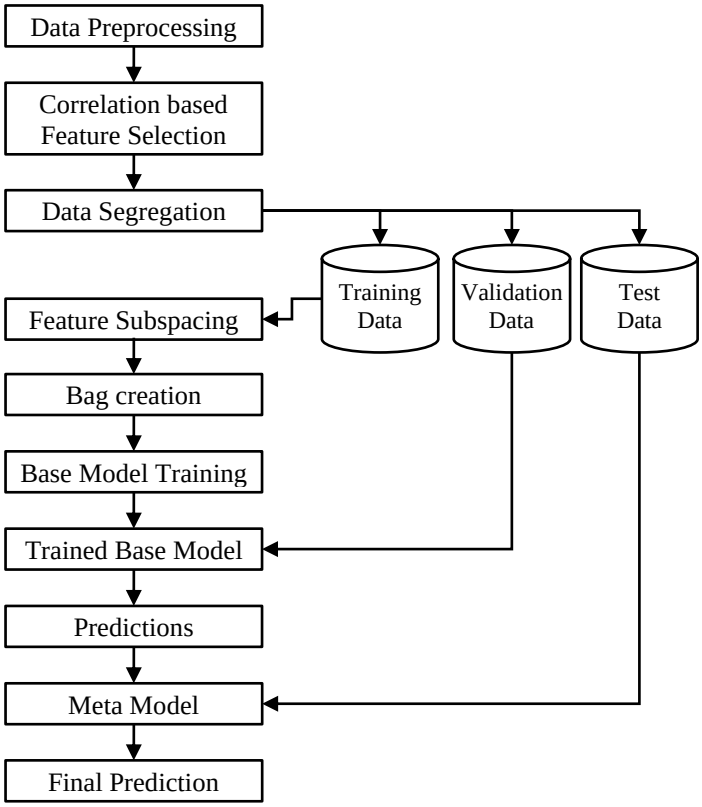

Fig.1. FSSM Architecture

The proposed stacking architecture is composed of two major phases; the initial phase performs feature sub-spacing and bag creation for the initial training process, while the second phase creates and trains meta-learner for final predictions. FSSM architecture is shown in Fig.1.

3.1 DATA PREPROCESSING

Data preprocessing is performed on the data to convert the data into processible format for the machine learning model. Data standardization is the major processes involved in this phase. The proposed FSSM model has been designed as a binary classifier. Hence, converting multi-class data into binary class is performed. This work uses Bank data and NSL KDD data. Class labels of Bank data is binary in nature, while that of NSL KDD is multi-class in nature. NSL KDD data is intrusion detection data. Hence, the class attributes represents normal traffic (class 0) and four different types of attacks. The attack classes are aggregated to form a single class and are labelled as1.

The next standardization process requires conversion of nominal data into numeric formats. Network data tends to contain nominal attributes depicting protocols and flags. Machine learning models are capable of operating on numeric data. Hence the nominal attributes are analyzed and are either converted or eliminated based on their status. String based attributes that represents IP addresses are eliminated, while nominal values such as protocols and transmission statues are one-hot-encoded. This marks the end of the preprocessing phase.

3.2 CORRELATION BASED FEATURE SELECTION

Feature selection plays a vital role in the machine learning process for a variety of reasons. Foremost, being the best solution to handle curse of dimensionality. Curse of dimensionality refers to the bias and complexity created in a machine learning model due to the presence of large number of attributes. Further, analysis shows that not all attributes contribute equally towards predictions. Some attributes tend to be neutral, while other exhibit positive or negative impact towards the prediction process. Neutral attributes and attributes that exhibit negative impacts should be eliminated. Feature selection is the process employed to identify such attributes. Correlation is one method that can be used to identify the significance of attributes. This work uses Pearson Correlation Coefficient (PCC) to determine the correlation level between all the attributes and the class attribute. Correlation between two attributes X and Y are given by,

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}} \quad (1)$$

Determining correlation between a data attribute and the class attribute reveals the relationship of the attribute towards the class. This technique is used to identify correlations between numerical attributes. Hence numerical encoding is performed prior to this process. Attributes that exhibit high correlation with the class attribute are selected. The threshold correlation level is determined by the domain expert after careful analysis of the data. The proposed work uses 30% as the threshold for bank data and 60% as the threshold for NSL KDD data.

3.3 FEATURE SUB-SPACING BASED BAG CREATION AND SEGREGATION

This is first phase of the data training process. Prior to the training process, the data is segregated into training, validation and test data. The training data is used for feature sub-spacing. Feature sub-spacing is the process of dividing and creating subsets of attributes in the data for the process of bag creation. Bags contain subsets of data for training. Usually instance based subset creation is performed. However, this work performs feature and instance based subset creation. The process of feature sub-spacing aims to reduce features in the bags, hence resulting in reduced dimensions, which in turn reduces the complexity levels of the data. Multiple data bags are created, each with varied set of features.

Multiple instances of machine learning models have been created. Each model is provided with its own data bag for training. Data division and feature sub-spacing results in reduced imbalance and reduced complexity levels. Every model is trained in different data. Hence decision rules obtained from each model varies considerably from other models. Decision Tree is used as the machine learning model of choice. The total number of models to be created is based on the complexity levels of the data. Highly complex data mandates several models, while data with low or moderate complexity will perform effectively on few models. This is decided by the domain expert. This forms the first phase of the stacking process, and the models created in this phase are called the base models.

3.4 STACKING BASED TRAINING AND PREDICTION

Stacking is the process of building models on top of existing models. This forms the second phase of the training process. Model in the second stacking phase is trained on the predictions obtained from the base models in the previous phase. The machine learning models used in this phase is known as the meta-model. The meta-model trains on the prediction process of the base models, rather than on the data. Validation data is passed to the initial phase, and the predictions obtained from the base models is integrated to form the training data for the stacking model. Class label for this data is obtained from the class label of the validation data. Logistic Regression is used as the meta-model. The meta-model harnesses the capabilities of the trained base models to provide predictions. Hence the quality of predictions obtained from the stacking model is higher compared to single level models.

The actual prediction process transpires by passing the test data to the trained base models. These models provide multiple predictions, which are integrated and passed to the meta-model. The meta-model predicts on the integrated data to provide the final prediction.

4. RESULTS AND DISCUSSION

The proposed FSSM model is analyzed on the Brazilian Bank data and NSL KDD data to identify the performance levels. Implementation has been performed in Python, and the obtained results were recorded.

4.1 PERFORMANCE ON BANK DATA

ROC plot, depicting the efficiency of the prediction model on Bank data is shown in Fig.2. ROC is plotted with False Positive Rate (FPR) in the x-axis and True Positive Rate (TPR) in the y-axis. Higher values of TPR pushes the plot towards the top, while low FPR levels pushes the plot towards the left. An ideal ROC plot has its peak on the top left, depicting high TPR and low FPR levels, and a large area under the curve. It could be observed from the Fig.2 that the proposed FSSM technique exhibits such a plot, depicting high prediction efficiency.

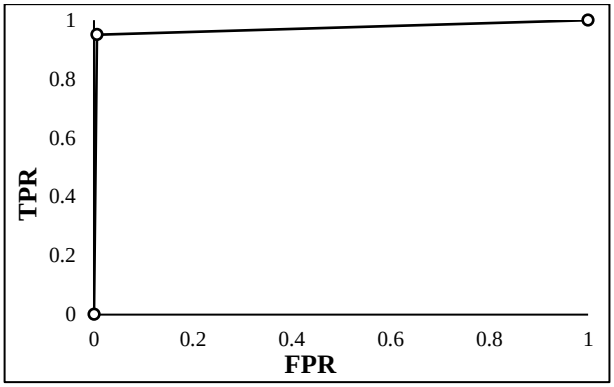

Fig.2. ROC Plot of FSSM

A view of the PR plot of the FSSM technique on Bank data is shown in Fig.3. PR plot is created with recall on the x-axis and precision on the y-axis. An effective model is required to exhibit high values of precision and recall. This pushes the peak of the curve towards the top left. It could be observed from the Fig.2 that the FSSM technique exhibits very high levels of precision and recall at >0.9. Hence the peak point of the plot is aligned towards the top right, depicting the prediction efficiency of the model.

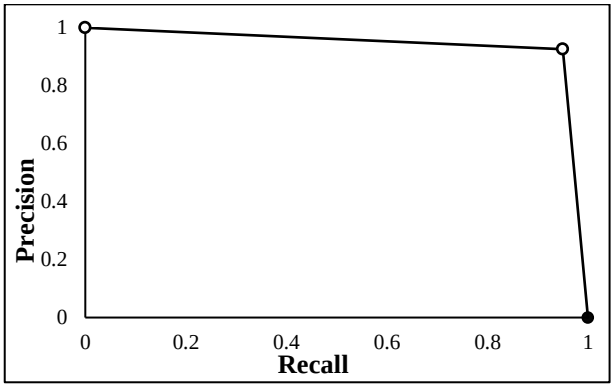

Fig.3. PR Plot of FSSM

4.2 PERFORMANCE ON NSL-KDD DATA

Performance on NSL-KDD data is shown in Table.1. NSL-KDD is highly imbalanced and multi-class in nature. The class values are converted to binary prior to analysis.

Table.1. Performance of FSSM on NSL-KDD

Metric	Value	Metric	Value
TPR	0.977842	Precision	0.941362
TNR	0.930593	Recall	0.977842

FPR	0.069407	Accuracy	0.955758
FNR	0.022158	F1-Score	0.959255

It could be observed that all the true prediction performances fare above 90%, while false predictions fall below 10% (FPR 6% and FNR 2%). This shows the capability of the FSSM model in operating on varied types of data with varied imbalance levels to provide highly effective results. Further, this also shows the generic nature of the model.

5. COMPARATIVE STUDY

The proposed FSSM technique is compared with the TWB model proposed by Somasundaram et al. [16] on performance from the Bank data. A comparison of the true prediction metrics, which corresponds to TPR and TNR are shown in Table.2. The proposed FSSM technique exhibits improved performance on both TPR and TNR levels. Improvements in TPR levels at 26% and TNR levels at 14% could be observed from the figure. The improvement in performance is attributed to the reduced complexity levels and also reduced imbalance levels, which are facilitated due to the initial preprocessing phases and the stacking model.

False prediction levels, depict FPR and FNR rates, and are shown in Table.2. False prediction levels are expected to be low in models exhibiting effective performances. False prediction levels of the proposed FSSM technique could be observed to be the lowest compared to the TWB model. This shows the high interpretability levels of the proposed FSSM model in correctly discriminating classes.

A tabulated comparison of FSSM with TWB is provided in Table.2. It could be observed that the FSSM model exhibits high performance in true prediction levels, false prediction levels and recall. A slight reduction of 3% has been observed in the precision levels, while all the other performances exhibit improvements.

Table.2. Performance Comparison of FSSM with TWB

Metrics	**FSSM**	**TWB**
TPR	0.9500	0.6932
FPR	0.0063	0.1455
TNR	0.9936	0.8544
FNR	0.0500	0.3067
Precision	0.9268	0.9530
Recall	0.9500	0.6932

6. CONCLUSION

Classification forms a major component of the current automated environment. Dependency on automation has increased the data collection levels, which has resulted in huge amounts of data for analysis. Classification is the process of analyzing data to obtain useful information. Imbalance contained in the real world data creates several complications in the classification process. This work proposes a generic classification architecture that can handle data imbalance effectively to provide unbiased predictions. The proposed Feature Sub-set Stacking Model (FSSM) has been designed as a binary classifier model that can handle highly imbalanced data. The FSSM technique is composed of two phases; the initial phase performs feature sub-spacing and base model training, and the second phase proposes a stacking architecture, which uses a meta-model to provide the final predictions. The major limitations of this model is that it is a binary classifier, and the precision levels of this model were found to be slightly lower than the existing models. Future enhancements will be based on extending the model to handle multi-class data and also measures to improve the precision levels.

REFERENCES

[1] A. Somasundaram and U.S. Reddy, "Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data", *Proceedings of International Conference on Research in Engineering, Computers and Technology*, pp. 1-16, 2016.

[2] A. Somasundaram and U.S. Reddy, "Modelling A Stable Classifier for Handling Large Scale Data with Noise and Imbalance", *Proceedings of International Conference on Computational Intelligence in Data Science*, pp. 1-6, 2017.

[3] M. Koziarski, B. Krawczyk and M. Woźniak, "Radial-Based Oversampling for Noisy Imbalanced Data Classification", *Neurocomputing*, Vol. 343, pp. 19-33, 2019.

[4] C. Tsai, W. Lin, Y. Hu and G. Yao, "Under-Sampling Class Imbalanced Datasets by Combining Clustering Analysis and Instance Selection", *Information Sciences*, Vol. 477, pp. 47-54, 2019.

[5] G. Chen and Z. Ge, "SVM-Tree and SVM-Forest Algorithms for Imbalanced Fault Classification in Industrial Processes", *IFAC Journal of Systems and Control*, Vol. 8, pp. 1-8, 2019.

[6] M. Buda, A. Maki and M.A. Mazurowski, "A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks", *Neural Networks*, Vol. 106, pp. 249-259, 2019.

[7] Y. Qian, Y. Liang, M. Li and G. Feng, "A Resampling Ensemble Algorithm for Classification of Imbalance Problems", *Neurocomputing*, Vol. 143, pp. 57-67, 2014.

[8] N.V. Chawla and K.W. Bowyer, "SMOTE: Synthetic Minority Over-Sampling Technique", Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357, 2002.

[9] X. Tao, Q. Li and W. Guo, "Self-Adaptive Cost Weights-Based Support Vector Machine Cost-Sensitive Ensemble for Imbalanced Data Classification", *Information Sciences*, Vol. 487, pp. 1-56, 2019.

[10] H. Faris, R. Abukhurma and W. Almanaseer, "Improving Financial Bankruptcy Prediction in a Highly Imbalanced Class Distribution using Oversampling and Ensemble Learning: A Case from the Spanish Market", *Progress in Artificial Intelligence*, Vol.9, No. 1, pp. 1-23, 2019.

[11] M. Lazaro, F. Herrera and A. Figueiras Vidal, "Ensembles of Cost-Diverse Bayesian Neural Learners for Imbalanced Binary Classification", *Information Sciences*, Vol. 520, pp. 31-45, 2020.

[12] X. Gao, "An Ensemble Imbalanced Classification Method based on Model Dynamic Selection Driven by Data Partition Hybrid Sampling", *Expert Systems with Applications*, Vol. 160, pp. 1-13, 2020.

[13] Y. Sui, Y. Wei and D. Zhao, "Computer-Aided Lung Nodule Recognition by SVM Classifier based on Combination of Random under Sampling and Smote", *Computational and Mathematical Methods in Medicine*, Vol. 620, pp. 1-13, 2015.

[14] Y. Zuo and C.Z. Jia, "Carsite: Identifying Carbonylated Sites of Human Proteins based on a One-Sided Selection Resampling method. Molecular Biosystems", Vol. 13, No. 11, pp. 2362-2369, 2017.

[15] P. Vuttipittayamongkol and E. Elyan, "Neighborhood-Based under Sampling Approach for Handling Imbalanced and Overlapped Data", *Information Sciences*, Vol. 509, pp. 47-70, 2020.

[16] A. Somasundaram and S. Reddy, "Parallel and Incremental Credit Card Fraud Detection Model to Handle Concept Drift and Data Imbalance", *Neural Computing and Applications*, Vol. 31, No. 1, pp. 3-14, 2018.

[17] S. Suh, H. Lee, P. Lukowicz and Y. Lee, "CEGAN: Classification Enhancement Generative Adversarial Networks for unraveling data imbalance problems", *Neural Networks*, Vol. 133, pp. 69-86, 2021.

[18] H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah and Q. Kang, "Optimizing Weighted Extreme Learning Machines for Imbalanced Classification and Application to Credit Card Fraud Detection", *Neurocomputing*, Vol. 407, pp. 50-62, 2020.

[19] Q.Y. Zhu, A.K. Qin, P.N. Suganthan and G.B. Huang, "Evolutionary Extreme Learning Machine", *Pattern Recognition*, Vol. 38, No. 10, pp. 1759-1763, 2005.

[20] J. Cao, Z. Lin and G. Huang, "Self-Adaptive Evolutionary Extreme Learning Machine", *Neural Processing Letters*, Vol. 36, No. 3, pp. 285-305, 2012.

[21] Y. Xu and Y. Shu, "Evolutionary Extreme Learning Machine - Based on Particle Swarm Optimization", *Proceedings of International Conference on Advances in Neural Networks*, pp. 1-26, 2006.

[22] F. Han, H. Yao and Q. Ling, "An Improved Evolutionary Extreme Learning Machine based on Particle Swarm Optimization", *Neurocomputing*, Vol. 116, pp. 87-93, 2013.

[23] C. Wang, C. Deng and S. Wang, "Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification with XGBoost", *Pattern Recognition Letters*, Vol. 136, pp. 190-197, 2020.

[24] H. He, W. Zhang and S. Zhang, "A Novel Ensemble Method for Credit Scoring: Adaption of Different Imbalance Ratios", *Expert System Applications*, Vol. 98, pp.105-117, 2018.

[25] Xu Ying Liu, Jianxin Wu and Zhi Hua Zhou, "Exploratory Undersampling for Class-Imbalance Learning", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 39, No. 2, pp. 539-550, 2009.

[26] M.F. Kabir and S. Ludwig, "Classification of Breast Cancer Risk Factors using Several Resampling Approaches", *Proceedings of IEEE International Conference on Machine Learning and Applications*, pp. 1243-1248, 2018.