Classification of Imbalanced Datasets using One-Class SVM, k-Nearest Neighbors and CART Algorithm

Maruthi Rohit Ayyagari
College of Business, University of Dallas
Irving, Texas, USA

***Abstract*—In this paper a new algorithm, OKC classifier is proposed that is a hybrid of One-Class SVM, k-Nearest Neighbours and CART algorithms. The performance of most of the classification algorithms is significantly influenced by certain characteristics of datasets on which these are modeled such as imbalance in class distribution, class overlapping, lack of density, etc. The proposed algorithm can perform the classification task on imbalanced datasets without re-sampling. This algorithm is compared against a few well known classification algorithms and on datasets having varying degrees of class imbalance and class overlap. The experimental results demonstrate that the proposed algorithm has performed better than a number of standard classification algorithms.**

Keywords*—SVM; k-NN; CART; OKC; classification; machine learning

I. Introduction

Classification is a task of categorizing the instances of a specified class from amongst the given set of classes. This task is done by a classifier that is demonstrated on a dataset of training cases. Most of the classification algorithms expect balanced class, i.e. there will be practically equivalent number of cases from all classes in the preparation dataset. But in many real world domains, like fraud detection, medical diagnosis, etc., the number of examples that belong to one class may severely outnumber the instances that belong to another class/classes. Such datasets, in which significant differences in the proportion of cases having a place with various classes are possible, called imbalanced datasets. The imbalance in class distribution could prompt high misclassification rates of minority class cases. One of the real explanations for this is the majority of the classification algorithms deal with the objective of enhancement of accuracy. As the majority class instances are much higher in number than the minority class ones, the classifier would give high accuracy, even if it classifies all instances as majority class and misclassifies all the minority class instances. This is called class imbalance problem. Besides the imbalanced datasets, other data intrinsic characteristics like overlapping between classes, presence of small disjuncts and lack of density of the minority class in training datasets could also impact the performance of the classifier significantly. The issue of class imbalance becomes more serious in the presence of one or more of such data on intrinsic characteristics. A few arrangements have been proposed in the past to manage these issues independently. In this paper, we have proposed a new algorithm, namely, OKC classifier (hybrid of One-class SVM, K-nearest neighbor and CART) to overcome this problem.

A. Imbalanced Datasets

In many real life applications, the situation of imbalanced datasets every now and again shows up. A dataset in which one class extremely outnumbers other can be considered as an imbalanced dataset. The class with moderately less number of cases in a dataset is called 'minority class' and alternate class is called ‘majority class’. The minority class usually represents the most essential idea to be learned, and it is hard to distinguish it since it may be related to huge and remarkable cases, or because the data acquisition of these cases is costly [1-2]. The imbalance of data distribution between different classes is known as between-class imbalance [3]. Such imbalance could be a consequence of the intrinsic nature of the data. For example, in the fraud detection domain, it is difficult to get the data related to the fraudulent transactions than the data that belong to legitimate transactions. Within a class imbalance is said to happen when a class is comprised of various sub-groups and the quantity of cases having a place with each sub-bunch is altogether not quite the same as those of other sub-bunches inside a similar class [4].

B. Class Overlapping

The class overlaps problem appears when a region in data space contains training data from more than one class. In such case, there is no clear partition between various classes causing difficulty in the classification process. The performance of a classifier is extraordinarily influenced when the issue of class overlapping is present along with an imbalance in the dataset. It has been proved that for the datasets that have clean clusters, i.e. no overlapping and are linearly separable, classifier performance on such datasets is not influenced by any degree of imbalance [1, 5]. In other works, it has been proved that if the data in the overlapping region are imbalanced, then the imbalance ratio affects the performance more than the size of overlap [1].

C. Lack of Density

The issue of lack of density emerges when there is almost no information accessible to represent the minority class concept. In the event that the cases of the minority class are less, then it becomes difficult to distinguish between minority class and noise. The majority of standard classifiers aim to

obtain a good generalization capability. In case of lack of density of a minority class, the classification rules that predict the minority class are highly specialized whereas due to the large number of majority class cases, the classification rules that predict the majority class seem to be more general to the classifier as their coverage is very high as compared to the minority class ones [6]. So, in this case the rules that predict the minority class are discarded by the classifier leading to high misclassification of a minority class.

II. Background

Verma et al. [7] used median filter, Gaussian filter and un-sharp masking J for the image enhancement. Entropy based segmentation is used to find the region of interest and then KNN and SVM classification techniques for the analysis of kidney stone images. The accuracy of KNN was found 89% and that of SVM was 84%. Li and Wang [8] used SIFT (Scale-invariant feature transform) algorithm to extract feature and the extracted features are clustered by K-means clustering algorithm. After clustering BOW (bag of word) of each image is constructed and multi-class classifier is trained using SVM (Support Vector Machine) to classify images. Authors reviled that SVM gave better results in small sample training set. Accuracy of image classification was about 90% with this method. Guo et al. [9] proposed SVM-based sequential classifier training (SCT-SVM) approach for remote sensing image classification. This technique help in reducing required number of training samples for classifier training. Different experiments were conducted with Sentinel-2A multitemporal data and accuracy of 76.18% to 94.02% achieved with the proposed technique.

McDermott et al. [10] in this study investigate Support Vector Machine (SVM) classifiers for detecting brain hemorrhages using Electrical Impedance Tomography (EIT) measurement frames. A 2-layer model of the head with series of hemorrhages is designed by means of numerical models and physical phantoms. Authors reported that phantom models are more challenging with maximum specificity of 75% when used with the linear SVM. The detection are was increased when radial basis function (RBF) SVM classifier and a neural network classifier were applied. Badgujar and Deore [11] proposed a hybrid algorithm using Migrating Bird Optimization and Support Vector Machine (MB-SVM) classifiers. Gaussian filters are used to eradicate the noise from the fundus retinal image. Experimental validation on a publicly available STARE data-set demonstrates the improved performance of the proposed method over existing method. Ma et al. [12]. Presented weight-KNN the KNN-based model acquire the test image's k-nearest neighbors and get the prediction of the image according to the contribution of its neighbors. Hu et al. [13] combine color, texture and shape feature towards multi-type feature. These features were integrated with k-nearest neighbor classifier. Experiment were conducted on 4500 aerial images and recognition rate of 99% was achieved using this multi-type feature. Gul et al. [14] propose an ensemble of subset of k-NN classifiers, (ESkNN) for classification. Experiments were conducted on benchmark data sets and results are compared with usual k-NN, bagged k-NN, random k-NN, multiple feature subset method, random forest and support vector machines. The proposed ensemble gives better classification performance than the usual k-NN and its ensembles, and performs comparable to random forest and support vector machines. Guo et al. [15] proposed a guided filter-based method and used two fusion methods for spectral and spatial features. Hyperspectral images were classified using SVM. The proposed method were fast in execution and easy to implement.

A. Proposed OKC Classifier

The proposed algorithm is a hybrid of one class SVM, k-Nearest Neighbour and CART (Classification and Regression Tree) algorithms. In this algorithm, Hellinger distance and Gini impurity are used as splitting criteria for choosing the best feature and best value to split, respectively. Hellinger distance has been proved to be skewed insensitive [16] i.e. it is not affected by the situation of class imbalance. On each leaf node of this tree where the illustrations have diverse classes, feature selection is done to choose two features that could best discriminate among the classes and then k-Nearest Neighbours is trained on all examples and one class SVM is trained on the minority class samples. When a new prediction is to be done, it is first classified to the leaf node and then it is categorized as inlier or outlier by the one class SVM. If it is predicted as inlier, it is assigned the minority class otherwise after feature selection it is assigned the class predicted by the k-Nearest Neighbor algorithm with k=1 i.e. the class of its nearest neighbour. This algorithm is designed to handle the class imbalance problem even if other data intrinsic characteristics like class overlap and lack of density is also present. As the feature selection is done at each leaf, only those features that play a significant role in classification are selected. It means that overlapping features will be discarded and thus the class overlapping problem can be handled to a great extent. The feature selection is done by using Hellinger distance [17]. The one class SVM algorithm is trained on the minority class tests at each leaf with mixed samples, so it is ensured that all minority class illustrations are learnt by the classifier.

B. One Class SVM

Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper-plane. The conventional 2-class classifier finds a hyper-plane that isolates one class from another. The one-class SVM finds the hyper-plane that separates all of the in-class points from the origin; it is essentially a two-class SVM where the origin is the only member of the second class. So, basically it separates all the data points from the origin and maximizes the distance from this hyper-plane to the origin. This results in a binary function, which captures regions in the input space and returns +1 in the region capturing the training data point & -1 elsewhere [18].

C. K-Nearest Neighbors

In the K-Nearest Neighbour algorithm, an object is classified by a majority vote of its neighbors, with the object being classified to the class most common among its k nearest

neighbors. If k=1, the object is simply classified to the class of that single nearest neighbour. It is typically in light of the Euclidean separation between a test sample and the specified training samples. For n-dimensional space, the Euclidean distance between two points x and y is calculated as following:-

$$d = \sqrt{\sum_{k=1}^{N}(x_k - y_k)^2}$$

It has been observed that the k-NN algorithm suffer from the curse of dimensionality [19] i.e. it cannot perform well when the number of features of the dataset is large. To deal with this issue, we are doing feature selection to select the best features that could best discriminate among the classes before applying k-NN. This feature selection is done at each leaf, with mixed class samples, separately so that the problem of class overlap could be minimized as different features may be prevalent in different places in the data space.

D. CART

Classification and Regression Tree (CART) is a binary recursive partitioning algorithm that is fit for handling nominal and continues attributes both as targets and predictors [20]. The classification tree is built by recursively splitting parent nodes into two child nodes that have maximum homogeneity. This homogeneity is determined by an impurity function. CART searches through all values of the attributes to find the best value to split. There are several impurity functions like Gini index, Towing splitting rule, etc. The process of splitting is stopped when a node becomes pure. Otherwise, it is repeated until a split result into a child node with less number of observations than a predefined number, or when the change in impurity function is less than the predefined minimum change number. Classification of a new observation is made by assigning the dominating class of the leaf node to which the new observation belongs to. In case of imbalanced datasets, when there is the problem of absolute rarity or lack of density of the minority class, the dominating class at the leaf nodes is usually the majority class. This results into misclassification of the observations that belong to the minority class. To sort out this problem, we are using the One-class SVM and k-NN at the leaf nodes with mixed classes instead of voting. One-class SVM is trained on the minority class, to cover all minority class examples so that the problem of lack of density of minority class can be handled to at least some extent. Then after selecting two features using Hellinger distance, k-NN is trained on all samples of the leaf.

E. Splitting Criteria for OKC Classifier

In the proposed algorithm, the splitting criteria used for the choice of the best features, is Hellinger distance and the criteria used for the selection of best value of the chosen feature is Gini impurity. Hellinger distance is a good criterion to be used with imbalanced datasets as it is not affected by the class distribution skew [16]. Assuming a binary class problem (class + and class-), let x+ be class+ and x- be class-, x_{+j} is the number of positives in bin j and x_{-j} is the number of negatives in bin j. For a feature that has p number of bins, the Hellinger distance is given below:

$$d_H(x_+, x_-) = \sqrt{\sum_{j=1}^{p}\left(\sqrt{\frac{|x_{+j}|}{|x_+|}} - \sqrt{\frac{|x_{-j}|}{|x_-|}}\right)^2}$$

The Hellinger distance for all features is calculated before each split and the feature with maximum Hellinger distance is chosen to split. After that, the choice of best value, of the selected feature, to split is made by using Gini impurity. Gini impurity is the expected error rate if one of the results from a set is randomly applied to one of the items in the set [20]. Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. To compute Gini Impurity for a set of items, suppose i {1, 2,…m}, and let f_i the fraction of the items labeled with the value i in the set, the Gini impurity as given below:-

$$I_G(f) = \sum_{i=1}^{m} f_i(i - f_i) = \sum_{i=1}^{m}(f_i - f_i^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2 = \sum_{i \neq k} f_i f_k$$

The value with the lowest Gini impurity is selected for split.

F. Stopping Conditions for OKC Classifier

The process of splitting of nodes is done recursively until some stopping condition is met. In the proposed algorithm, there are three stopping conditions:

1) When the node becomes pure i.e. all samples on that node belongs to a single class.

2) If the change in impurity functions, i.e. Gini index, after splitting, would be less than the predefined minimum value.

3) If the split would result into a child node with less number of samples than the predefined minimum number of samples.

G. Algorithm

Input: A set S of labeled instances, threshold values for minimum number of samples at leaves and minimum change in utility function i.e. Gini Impurity.

Output: A binary tree with class labels and/or one class SVM, list of selected features and k-NN classifiers at leaves.

Step 1 If all samples at the current node have the same labels, assign that label to the current node and return.

Step 2 For each attribute, evaluate the hellinger distance and choose the attribute A with a maximum value of the hellinger distance.

Step 3 For each distinct value of A, evaluate Gini impurity and choose the value V, with the lowest value of Gini Impurity.

Step 4 Evaluate the difference between the utility of the current node and the utility that would result after split is performed on value V of attribute A.

Step 5 If the difference in utility is less than the threshold value or if the split would result into nodes with the less number of samples than the threshold value, fit a one-class SVM on the minority class samples and calculate hellinger distance on all attributes to choose two attributes with highest and the second highest value of hellinger distance. On the chosen attributes fit a k-NN classifier and return.

Step 6 Partition S with value V and attribute A.

For each child node, call the algorithm recursively.

III. Evaluation and Discussion

In this work, we have considered public dataset of five categories, namely, Yeast, CTG, Wilt, Fraud, and Semiconductors. Brief information about these databases also depicted in Table I.

A. Experimental Results

In order to evaluate the performance of the proposed algorithm, we have considered five different public datasets as described in Section 3. These five different datasets are normalized and taken from the UCI repository [21]. The results of the proposed algorithm are compared with standard machine learning algorithms decision tree, neural network, SVM, Naïve Bayes, k-Nearest Neighbors, Naive Bayes tree and CART. The proposed algorithm is also compared against random over-sampling, random under sampling, hybrid over-under sampling and meta-cost techniques applied to all the standard algorithms discussed in this section. In meta-cost, the cost of misclassification of minority class is set to double than the cost of misclassification of the majority class. The results obtained after performing various experiments without sampling, after random under-sampling and after random over-sampling are depicted in Tables II to IV, respectively. Experimental results based on hybrid of random under-sampling and random over-sampling are presented in Table V. In Table VI, we have presented experimental results achieved after setting meta-cost double for misclassification of minority class than the misclassification of the majority class. We have seen that proposed classification algorithm, namely, OKC performs better than existing algorithms.

TABLE I. Datasets used for Experiments

Dataset	Number of Attributes	Size of Training Data		Size of Testing Data	
		Class-I	Class-II	Class-I	Class-II
Yeast	9	20	464	10	199
CTG	22	365	1124	106	531
Wilt	6	74	4265	187	313
Fraud	25	100	600	200	100
Semiconductor	3	76	924	28	539

TABLE II. F-Score (%age) without Sampling

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	41.38	0	0	0	0
Neural Network	51.41	14.85	18.18	28.33	0
SVM	55.06	0	0	0	0
Naïve Bayes	41.03	3.16	8.42	34.68	0
k-NN	11.64	0	0	7.69	0
Naïve-Bayes Tree	22.47	15.76	0	23.38	0
CART	21.95	4.19	0	0	0
Proposed Algorithm	61.27	59.23	22.9	80.7	21.05

TABLE III. F-Score (%age) under Sampling

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	22.1	41.41	9.04	63.8	9.24
Neural Network	43.24	35.65	9	61.4	7.9
SVM	49.83	20.85	8.99	57.62	6.4
Naïve Bayes	49.04	40.33	8.7	65.59	7.35
k-NN	19.59	55.42	9.57	59.42	7.35
Naïve-Bayes Tree	46.46	37.68	8.29	69.57	0
CART	36.96	44.11	9.04	69.32	8.7
Proposed Algorithm	61.27	59.23	22.9	80.7	21.05

TABLE IV. F-Score (%age) over Sampling

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	41.84	28.05	10.22	79.41	8.81
Neural Network	49.83	26.05	7.92	36.65	9.9
SVM	52.17	7.14	8.75	65	7.16
Naïve Bayes	40.82	45.85	8.7	72.05	7.23
k-NN	22.5	23.26	10.53	54.42	12.95
Naïve-Bayes Tree	21.43	10.1	8	30.95	8.76
CART	44.02	29.46	8.22	46.59	12.77
Proposed Algorithm	61.27	59.23	22.9	80.7	21.05

TABLE V. F-Score (%age) after Hybrid Under-Sampling and Over-Sampling

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	32.21	37.66	10.69	61.16	9.51
Neural Network	51.27	19.23	5.37	50.87	9.63
SVM	50.67	11.94	7.82	60.32	6.73
Naïve Bayes	45.42	45.96	7.95	64.42	7.43
k-NN	22.32	57.33	8.39	56.19	10.77
Naïve-Bayes Tree	27.35	33.62	6.50	52.20	6.45
CART	51.33	50.97	14.46	53.64	9.21
Proposed Algorithm	61.27	59.23	22.90	80.70	21.05

TABLE VI. F-SCORE (%AGE) AFTER SETTING META-COST

Algorithm	CTG	Wilt	Yeast	Fraud	Semiconductor
Decision Tree	14.94	0	0	8.92	0
Neural Network	50.21	22.75	0	51.45	0
SVM	52.31	0	9.31	0	0
Naïve Bayes	49.8	15.32	0	36.84	0
k-NN	27.03	7.14	0	12.88	3.57
Naïve-Bayes Tree	25.47	4.17	20	22.16	0
CART	20.65	5.18	0	45.54	0
Proposed Algorithm	61.27	59.23	22.9	80.7	21.05

IV. CONCLUSION

In this paper, a new classification algorithm based on a hybrid combination of one class SVM, k-NN and CART algorithms has been proposed. This algorithm is outlined to such an extent that it could perform well in classification of imbalanced datasets that are non-linearly separable without any need of resampling. Also, it can deal with the circumstances of class overlap and lack of density of the minority class in imbalanced datasets. Our experiments have shown that the proposed algorithm could outperform a number of standard classification algorithms. However, this work is focused only on the binary classification tasks. The task of multiclass classification in the presence of class overlaps, lack of density of the minority class in imbalanced datasets is left for future scope.

REFERENCES

[1] Asuncion A and Newman D (2007) UCI machine learning repository. http://archive.ics.uci.edu/ml/datasets.html.

[2] Cieslak DA, Hoens TR, Chawla NV, and Kegelmeyer WP (2012) Hellinger distance decision trees are robust and skew-insensitive. Data Mining and Knowledge Discovery, 24(1):136-158.

[3] Haibo He and Garcia E (2009) Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9):1263-1284.

[4] Japkowicz N (2001) Concept-Learning in the presence of Between-class and Within-class Imbalances. Advances in Artificial Intelligence, 67-77.

[5] Kouiroukidis N and Evangelidis G (2011) The effects of dimensionality curse in high dimensional k-nn search. In the proceedings of the 15th Panhellenic Conference on Informatics, 41-45.

[6] López V, Fernández A, García S, Palade V, and Herrera F (2013) An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250:113-141.

[7] Verma J, Nath M, Tripathi P and Saini KK (2017) Analysis and identification of kidney stone using Kthnearest neighbour (KNN) and support vector machine (SVM) classification techniques. Pattern Recognition and Image Analysis, 27:574. https://doi.org/10.1134/S1054661817030294.

[8] Li Q and Wang X (2018) Image Classification Based on SIFT and SVM. IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 762-765.

[9] Guo Y, Yin X, Zhao X, Yang D and Bai Y (2019) Wireless Com Network. EURASIP Journal on Wireless Communications and Networking. https://doi.org/10.1186/s13638-019-1346-z.

[10] McDermott B, O'Halloran M, Porter E, Santorelli A (2018) Brain haemorrhage detection using a SVM classifier with electrical impedance tomography measurement frames. PLoS ONE 13(7):e0200469. https://doi.org/10.1371/journal.pone.0200469.

[11] Badgujar R and Deore P (2018) MBO-SVM-based exudate classification in fundus retinal images of diabetic patients. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 1–12.

[12] Ma Y, Xie Q, Liu Y and Xion S (2019) A weighted KNN-based automatic image annotation method. Neural Computing and Applications. https://doi.org/10.1007/s00521-019-04114-y.

[13] Hu G, Yang Z, Zhu M, Li H, and Xiong N (2018) Wireless Com Network. https://doi.org/10.1186/s13638-018-1195-1.

[14] Gul A, Perperoglou A, and Khan Z (2018) Osama Mahmoud Miftahuddin Miftahuddin Werner Adler Berthold Lausen. Advanced Data Analysis and Classification, 12: 827. https://doi.org/10.1007/s11634-015-0227-5.

[15] Guo Y, Jia X and Paull S (2018) Effective Sequential Classifier Training for SVM-Based Multi-temporal Remote Sensing Image Classification, IEEE Transactions on Image Processing, 27(6):3036-3048.

[16] Luengo J, Fernández A, García S, and Herrera F (2011) Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary under sampling. Soft Computing, 15(10):1909-1936.

[17] Prati RC, Batista GE, Monard MC (2004) Class imbalances versus class overlapping: an analysis of a learning system behavior. Advances in Artificial Intelligence, 312-321.

[18] Schölkopf B, Williamson R, Smola A, and Shawe J (1999) Support Vector Method for Novelty Detection. In the proceedings of the 12th International Conference on Neural Information Processing Systems, 12:582-588.

[19] Seragan T (2007) Programming collective intelligence: building smart web2.0 application.

[20] Weiss GM (2005) Mining with rare cases. The Data Mining and Knowledge Discovery Handbook, Springer, 765–776.

[21] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, and McLachlan GJ (2008) Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1):1-37.