NEAREST-INSTANCE-CENTROID-ESTIMATION LINEAR DISCRIMINANT ANALYSIS (NICE LDA)

Rishabh Singh, Kan Li (Member, IEEE) and Jose C. Principe (Fellow, IEEE)

University of Florida
Department of Electrical and Computer Engineering

ABSTRACT

We propose a novel cascaded classification technique called the Nearest Instance Centroid Estimation (NICE) LDA algorithm. Our algorithm (inspired from NICE KLMS) performs a cascade combination of two weak classifiers - threshold based class-wise clustering and linear discriminant classification to achieve state-of-the-art results on various high dimensional UCI datasets. We show how our method is more robust towards skewed data and computationally more efficient than previous methods of combining clustering with classification techniques. We also develop an efficient aggregation method based on instance based learning that implements this cascade combination of classifiers in a much simpler manner computationally. We demonstrate that our method of data clustering and LDA implementation, while introducing only one free parameter, leads to results that are similar and often better than those achieved by the state-of-the-art kernel RBF SVMs.

Index Terms— LDA, Clustering, SVM, Cascade, Classification, UCI

1. INTRODUCTION

Combined classification methods use multiple weak classifiers in a logical way typically determined by the properties of the problem space to obtain good classification results typically in highly non-linear data. A single non-linear classifier, in such cases, may not be sufficient or may prove to be highly computationally intensive most of the times. A popular family of combined classifiers is ensemble classification techniques where the results of multiple classifiers are combined together in various ways to get the overall result. Polikar has given a detailed summary of ensemble based systems in [1] and stresses upon the fact that the two main components associated with ensemble classifiers are the diversity of individual classification techniques within the ensemble and the need to effectively combine the outputs of each individual classifier.

The paper is organized as follows. Section II gives a summary of the related work in the field of ensemble classification techniques. The recently introduced NICE KLMS learning scheme for kernel adaptive filters has been described. Section III introduces NICE LDA method and discusses some of its key features. Section IV discusses simulation results of NICE LDA implementation on some high dimensional 2 channel non-linear UCI datasets. We conclude in section IV and discuss some future scope related to this work.

2. BACKGROUND INFORMATION

2.1. Related Work

Over the past three decades, various different ensemble techniques have been developed in an effort to make it easier to deal with very large, high dimensional and highly non-linear data. The significant ways in which these methods vary from each other is in the techniques they use for partitioning the input data or for combining outputs from multiple models. Breiman in [2] presents a simple and effective bagging (short for bootstrap aggregation) predictor where the final output is determined from the average of the outputs of various versions of the predictor or from the majority vote of the multiple predictor versions. [3] presents a variant of this method called Random Forests which involves decision trees. Schapire [4] presents the Boosting algorithm which vastly improves the performance of a weak algorithm by using a learning algorithm within the model to eventually make its predictions accurate. This is done by making use of the prior knowledge of the weak learner's performance. This method however, does not work for a small number of samples. Freund and Schapire [5] later on present another method called AdaBoost that significantly improves upon the previous Boosting method by requiring no prior knowledge of the weak learner's performance to improve it. Wolpert in [6] presents a two level ensemble classifier system (stacked generalization) where the output of one ensemble of classifiers is fed to the input of another ensemble of classifiers which in turn attempts to learn the relationship between the output of the first ensemble and the ground truth values. Mixture of local experts is another algorithm based on the same concept [7].

Ensemble methods based on SVMs for classification have been used extensively for classification, especially in the field of computer vision. Authors in [8] present an ensemble of exemplar SVMs for object detection where a linear SVM is trained for each exemplar in the training set. This strategy of

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

training SVMs in a one-vs-rest sense has proven to be effective in their application in object detection, especially proving to be good for generalization. However, SVMs face major drawbacks due to their high computational complexity and their inefficiency in dealing with unbalanced data [9]. They get easily skewed towards the smaller class when the data is unbalanced and also become more sensitive to noise [10]. Authors in [9] attempt to overcome these drawbacks of ensemble SVMs by introducing a hybrid classifier that assigns a prior to the negative samples in a binary dataset and finds a hyperplane that separates this prior from the positive class.
Focus has been shifting towards LDA [11] based ensemble classifiers which have shown their performance to be similar to that of ensemble SVMs [12], [13]. In a typical ensemble LDA model, data is partitioned or clustered into various parts and an LDA classifier is trained for each part. A key difference between LDA and SVM is that an LDA gives equal importance to all the samples of the data unlike SVMs which tend to rely more on the marginal samples to construct the boundary [12]. Authors in [12] demonstrate how multiple LDAs with boosting and nearest neighbors form a powerful ensemble technique. Other variants of ensemble LDAs are shown in [14] and [15].
All of these methods of ensemble LDA classifiers perform clustering on the entire dataset during training without discriminating between classes. This could lead to very unbalanced partitions or clusters with very few data samples which is known to adversely affect the performance of LDA classifiers [16]. They use K-means as the clustering algorithm where they have to choose the number of clusters heuristically thus introducing a free parameter (number of clusters) that could vary drastically from one dataset to another. Furthermore, the aggregation methods used in all of these algorithms have to incorporate the decisions of all of the pretrained LDAs using majority vote, weighted majority vote or other similar techniques. A basic problem with majority vote is that the decisions tend to be biased if the class distribution is skewed. Authors in [13] attempt to overcome the problem of possible lack of data in the partitions or clusters by using an ensemble classifier based on regularized LDA (RDA). RDA evaluation requires that we optimize two different parameters making it very complicated. Our method of cascading clustering and implementing LDA offers a computationally simpler approach and overcomes all of these drawbacks associated with the existing ensemble models.

2.2. NICE KLMS

The Nearest Instance Centroid Estimation KLMS algorithm was recently introduced to restrict the growth of radial basis function structure and enable transfer learning in kernel adaptive algorithms [17]. It achieves this by using a group of local supports (data samples) in the input/feature space to perform kernel function approximation instead of using all of the samples (as is done in conventional kernel adaptive filtering). A Content Addressable Filter Bank (CAFB) is formed consisting of filters corresponding to each of the groups of supports in the input space. The groups of local supports are updated online iteratively by following the nearest-neighbors approach and using an appropriate centroid distance threshold. Hence, consequently the filter bank is also always being updated online and the appropriate filter is chosen for function approximation corresponding to the group of local supports that the new data point is closest to.

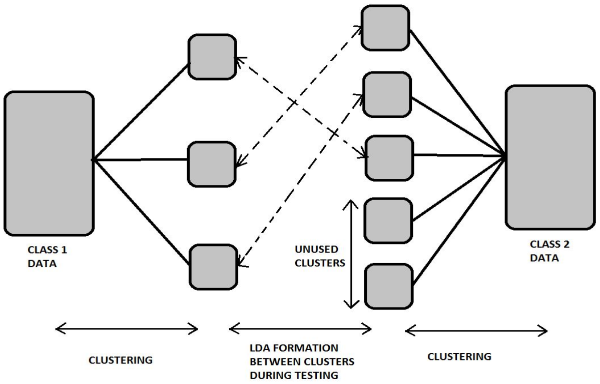

Fig. 1: NICE LDA: We implement clustering in the training phase and LDA models are trained instantaneously while testing between the cluster associated with the test data and the closest cluster in the opposite class

In other words, NICE KLMS is the decomposed approximate (because of finite arithmetic) orthogonal sum representation of the kernel function approximation over the entire input space given by:

$$f(x) = \sum_{i=1}^{N_1} \alpha_i^1 \phi(x_i, x) + ... + \sum_{i=1}^{N_p} \alpha_i^p \phi(x_i, x) \tag{1}$$

where each summation term represents a filter in the CAFB defined in an input space partition with N_t samples.

3. NICE LDA

We propose a classification technique consisting of a cascaded combination (unlike parallel combination used in ensemble methods) of nearest-neighbors based clustering and linear discriminant analysis. We call it the Nearest Instance Centroid Estimation (NICE) LDA. Our algorithm achieves in classification problem what NICE KLMS achieves in regression. In this case, we are replacing a computationally intensive kernel SVM algorithm (typically used in achieving the best results) with a content addressable bank of LDA models to achieve similar results. An important aspect of our algorithm is that we perform clustering on each class separately and customized to the data distribution of the particular class during training. We then use an efficient aggregation method during testing to form an iteratively updated bank of LDA models between pairs of opposite classes' clusters. The closest cluster to the test point is chosen as the first cluster. The cluster in the class opposite to that of the first cluster and nearest to it, is chosen as the second cluster with which the LDA model is to be formed. This LDA model finally projects the test point thereby classifying it. The algorithm is depicted in Fig. 1. Following are the key features of our algorithm:

- We first perform clustering on each class of the dataset separately using a distance threshold. If the distance of a training sample to its nearest cluster does not exceed this threshold, it is incorporated into the cluster and the cluster centroid is updated. Otherwise, a new cluster is declared using that training sample as its centroid. In each class, this distance threshold is set as a multiple of the standard deviation of the data thus performing clustering customized to the data distribution of the class and thereby minimize skewed partitions. Therefore, to a certain extent, we mitigate the adverse effects of unbalanced data on LDA performance mentioned in [16].
- We largely ensure adequate number of samples in each cluster through efficient class-wise clustering as described above. However, we may still encounter clusters where the number of samples is comparable to the number of data dimensions in which case the sample covariance matrix obtained through conventional LDA formation for those clusters may become singular. In such cases, we use an approximation of the covariance matrix and consider it only along its diagonal leading to a new class discrimination rule [18]. This is known as diagonal LDA and can be simply understood as a form of naive Bayes classification. In cases involving very high dimensional data, diagonal LDA is often a more appropriate variant of LDA for classification and is also much simpler than RDA.
- We only introduce one free parameter which is the common factor by which the standard deviation in each class is to be multiplied with to set the distance threshold for clustering in the respective class. Furthermore, we are using only one LDA model for test point classification unlike ensemble methods that typically use the majority vote of all classifiers connected in parallel to make decisions.
- We develop an instance based approach of updating the LDA bank where formation of LDA models take place only during testing when the nearest cluster to the test point is determined. We store the iteratively formed LDA models in a dictionary/bank during testing and refer to them whenever we encounter a test point that is close to a stored LDA's associated cluster. This way we ensure that no computation is wasted by forming LDAs associated with unused clusters.

Algorithm 1 NICE LDA

```
Initialization:
Std_1:    Class 1 standard deviation
Std_2:    Class 2 standard deviation
D_th:     Common Threshold Multiplier
$D_1 = Std_1 * D_{th}$:   Class 1 distance threshold
$D_2 = Std_2 * D_{th}$:   Class 2 distance threshold
Training (Clustering):
for all $m \in Classes$ do
    $C_{1,m} = x_{1,m}$: First training sample as cluster 1 centroid
    $C_m = [C_{1,m}]$: Cluster Dictionary
    while $x_{i,m} \in N_{train}$ do
        Closest centroid distance
        $d_{min} = \min_{1 \leq j \leq |C_m|} \|x_{i,m} - C_{j,m}\|^2$
        Nearest cluster
        $j* = \arg\min_{1 \leq j \leq |C_m|} \|x_{i,m} - C_{j,m}\|^2$
        if $d_{min} < D_m$ then
            Update Cluster j* centroid c and size s
            $c_{j*,m} = \frac{s_{j*,m} \cdot c_{j*,m} + x_{i,m}}{s_{j*,m} + 1}$
            $s_{j*,m} = s_{j*,m} + 1$
        else
            New Cluster Formation
            $C_{|C|+1,m} = []$: new cluster in class m
            $c_{|C|+1,m} = x_{i,m}$: new cluster centroid
            $s_{|C|+1,m} = 1$: effective size of cluster
            $C_m = [C_m, C_{|C|+1,m}]$
Testing:
$LD = [\quad]$: LDA dictionary
$C = [C_1 \ldots C_{|Classes|}]$
while $x_i \in N_{test}$ do
    Nearest neighbor cluster
    $j* = \arg\min_{1 \leq j \leq |C|} \|x_i - C_j\|^2$
    if $j* \in Class(P)$ then
        Find closest cluster of other class
        while $k \notin p$ do
            $g* = \arg\min_k \|C_k - C_{j*}\|^2$
        LDA training
        if $L_{j*,g*} \notin LD$ then
            $L_{j*,g*} = LDA(C_{g*}, C_{j*})$
            $L_{j*,g*}(x_i)$: Classification
            $LD = [LD, L_{j*,g*}]$: Update LDA dictionary
        else
            $L_{j*,g*}(x_i)$: Classification
```

CLASSIFICATION ACCURACY							
Algorithm	Ionosphere Data	Parkinsons Disease Data	Cleveland Heart Disease	Cardiac SPECT Data	Wisconsin Breast Cancer	Australian Credit Log	PIMA Indian Diabetes
SVM RBF	96.0265	92.6316	80.4124	78.6096	97.3085	86.1538	80.3419
CLUSTERING	90.7285	89.4737	76.2887	74.3316	96.8944	64.87	71.42
NICE LDA	94.0397	91.5789	84.5361	78.0749	97.7014	85.89	78.2

Table 1: Classification accuracy obtained from different methods: The class-wise clustering process before the LDA implementation alone achieves good accuracy values.

4. EXPERIMENTAL RESULTS

We evaluated the performance of the proposed NICE LDA on 7 popular UCI datasets of varying dimensionality and non-linearity and compared it with the best results obtained from the kernel based RBF SVM classifier on the same datasets. We used MATLAB 2017a for all simulations. Our training sample size was close to 50% of the total data samples. The results are shown in table 1.

It can be seen that the results of NICE LDA are very similar and in some instances even better than the best results obtained from kernel based RBF SVM even after keeping a low training size to dimensionality ratio. It is also seen here that our method of clustering alone accomplishes a big chunk of the classification task while the LDA classifiers contribute further to the accuracy to finally achieve state-of-the-art results.

Table 2 shows the parameters of the best performing NICE LDA models and details of the dataset. We notice that there is a healthy number and proportion of clusters formed in the both the classes which corresponds to the proportion of data samples in each class. This way, skewness of partitions is avoided. We also see that the best threshold value (the common factor by which the standard deviation of each class is multiplied) is a significantly large range of values for each UCI dataset.

5. CONCLUSION

In this paper, we demonstrated a new cascade classification scheme consisting of clustering followed by LDA implementation where we focused on class-wise clustering of data and utilized the information given by the standard deviation of data in each class. We are hence able to achieve a more balanced partitioning of data. This enables us to reliably use the LDA classifier and its diagonal variant. We are able to achieve classification results similar to the best results obtained by kernel based RBF SVMs on several high dimensional and non-linear UCI datasets. We only introduce one free parameter in our method and for each test sample we only rely on the output of one LDA model instead of using majority vote based techniques. We also develop an efficient instance based learning technique of forming the LDAs only during testing thereby preventing wastage of computation. This cascaded classification scheme along with its aggregation method can be potentially useful for classification of large online streaming data. We intend to explore this application in the future.

NICE LDA PARAMETERS							
	Ionosphere Data	Parkinsons Disease Data	Cleveland Heart Disease	Cardiac SPECT Data	Wisconsin Breast Cancer	Australian Credit Log	PIMA Indian Diabetes
Best Threshold Multiplier	0.65 - 0.77	0.64 - 0.69	1.41 - 1.57	1.3 - 1.32	1.55 - 1.67	1.26 - 1.42	1.53 - 1.73
Class 1:Class 2 (no. of Clusters)	9 : 6	12 : 6	3 : 2	2 : 2	1 : 2	3 : 3	4 : 4
Class 1:Class 2 (no. of Samples)	101 : 99	73 : 27	110 : 90	40 : 40	86 : 114	133 : 167	114 : 186
Train : Test (no. of Samples)	200 : 151	100 : 95	200 : 97	80 : 187	200 : 483	300 : 390	300 : 468
No. of Features	33	22	13	22	9	14	8

Table 2: NICE-LDA parameters: The number of clusters in each class is proportional to sample size of the class. Threshold multiplier values for which best classification performance is obtained has a significant range.

6. REFERENCES

[1] Robi Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine, vol. 6, no. 3*, 2006.

[2] L. Breiman, "Bagging predictors," *Machine Learning, vol. 24, no. 2*, 1996.

[3] L. Breiman, "Random forests," *Machine Learning, vol. 45, no. 1*, 2001.

[4] R.E.Schapire, "The strength of weak learnability," *Machine Learning, vol. 5, no. 2*, 1990.

[5] Y. Freund, and R.E.Schapire, "Decision-theoretic generalization to online learning and an application to boosting," *Journal of Computer and System Sciences, vol. 55, no. 1*, 1997.

[6] D.H.Wolpert, "Stacked generalization," *Neural Networks, vol. 5, no. 2*, 1992.

[7] R.A. Jacobs, M.L. Jordan, S.J. Nowlan, and G.E. Hinton, "Adaptive mixture of local experts," *Neural Computations, vol. 3*, 1991.

[8] Tomasz Malisievicz, Abhinav Gupta, and Alexei A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *International Conference on Computer Vision*, 2011.

[9] Margarita Osadchy, Daniel Keren, and Dolev Raviv, "Recognition using hybrid classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 4*, 2016.

[10] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *15th European Conference on Machine Learning*, 2004.

[11] R.A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugene*, 1936.

[12] M. Deypir, R. Boostani, and T. Zoughi, "Ensemble based multi-linear discriminant analysis with boosting and nearest neighbor," *Scientia Iranica, vol. 19, no. 3*, 2012.

[13] Akinari Onishi, and Kiyohisa Natsume, "Ensemble regularized linear discriminant analysis classifier for p300-based brain-computer interface," in *35th Annual International Conference of the IEEE EMBS*, 2013.

[14] X. Liu, L. Zhang, M. Li, H. Zhang, and D. Wang, "Boosting image classification with lda-based feature combination for digital photograph management," *Pattern Recognition, vol. 38, no. 6*, 2005.

[15] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, and S.Z. Li, "Ensemble-based discriminant learning with boosting for face recognition," *IEEE Transactions on Neural Networks, vol. 17, no. 1*, 2006.

[16] Jigang Xie, and Zhengding Qiu, "The effect of imbalanced data sets on lda: A theoretical and empirical analysis," *Pattern Recognition, vol. 40*, 2007.

[17] Kan Li, and Jose Principe, "Transfer learning in adaptive filters: The nearest-instance-centroid-estimation kernel least-mean-square algorithm," *IEEE Transactions on Signal Processing, vol. 65, no. 24*, 2017.

[18] Peter J. Bickel, and Elizaveta Levina, "Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations," *Bernoulli Society for Mathematical Statistics and Probability, vol. 10, no. 6*, 2004.