University of Dayton

eCommons

Electrical and Computer Engineering Faculty Publications

Department of Electrical and Computer Engineering

12-3-2018

Performance Analysis of Feature Selection Techniques for Support Vector Machine and its Application for Lung Nodule Detection

Barath Narayanan
University of Dayton

Russell C. Hardie
University of Dayton, rhardie1@udayton.edu

Temesguen Messay Kebede
University of Dayton

Follow this and additional works at: https://ecommons.udayton.edu/ece_fac_pub

 Part of the Electrical and Computer Engineering Commons

eCommons Citation

Narayanan, Barath; Hardie, Russell C.; and Kebede, Temesguen Messay, "Performance Analysis of Feature Selection Techniques for Support Vector Machine and its Application for Lung Nodule Detection" (2018). *Electrical and Computer Engineering Faculty Publications*. 435.
https://ecommons.udayton.edu/ece_fac_pub/435

This Conference Paper is brought to you for free and open access by the Department of Electrical and Computer Engineering at eCommons. It has been accepted for inclusion in Electrical and Computer Engineering Faculty Publications by an authorized administrator of eCommons. For more information, please contact mschlangen1@udayton.edu, ecommons@udayton.edu.

Performance Analysis of Feature Selection Techniques for Support Vector Machine and its Application for Lung Nodule Detection

Barath Narayanan Narayanan, Russell C. Hardie, Temesguen M. Kebede
Department of Electrical and Computer Engineering
University of Dayton, Dayton, OH 45469, USA
Email: {narayananb1, rhardie1, tmessay1}@udayton.edu

***Abstract* - Lung cancer typically exhibits its presence with the formation of pulmonary nodules. Computer Aided Detection (CAD) of such nodules in CT scans would be of valuable help in lung cancer screening. Typical CAD system is comprised of a candidate detector and a feature-based classifier. In this research, we study and explore the performance of Support Vector Machine (SVM) based on a large set of features. We study the performance of SVM as a function of the number of features. Our results indicate that SVM is more robust and computationally faster with a large set of features and less prone to over-training when compared to traditional classifiers. In addition, we also present a computationally efficient approach for selecting features for SVM. Results are presented for a publicly available Lung Nodule Analysis 2016 dataset. Our results based on 10-fold validation indicate that SVM based classification method outperforms the fisher linear discriminant classifier by 14.8%.**

***Index Terms* – Computer Aided Detection, Support Vector Machine, Computed Tomography, Lung Nodule, Fischer Linear Discriminant Classifier.**

I. INTRODUCTION

LUNG cancer is the leading cause of cancer death in the United States. 234,030 lung and bronchus cancer new cases are expected by the end of year 2018 [1]. 154,050 lung cancer deaths are expected by the end of the year 2018 [1]. Lung cancer typically exhibits its presence with the formation of pulmonary nodules. Early detection of such potentially cancerous nodules could improve patients' chances of survival [2]. Nodules are ellipsoidal growth present in the lung. Computed Tomography (CT) scans have proven to be effective for lung cancer screening in the past decade [2] and are currently employed by radiologists to detect such nodules. CT provides numerous slices of image data which can be time consuming and potentially fatiguing for radiologists to study. Hence, a Computer Aided Detection (CAD) system to automatically detect pulmonary nodules would be valuable for lung cancer screening and would enhance the workflow of a radiologist.

CAD of lung nodules has been a research area attracting great interest for the last few decades. Several CAD research papers have been presented in the literature [2-19]. In [3], a CAD system to detect lung nodules in CT scans is presented. Potential candidates are segmented and detected simultaneously using morphological operations. Later, a Fisher Linear Discriminant (FLD) classifier is utilized to classify the candidates based on a large suite of features. In [4], optimized method of feature selection is implemented for both clustering and classification using Sequential Forward Selection (SFS) for better CAD performance in CT scans and chest radiographs. In [5], a CAD system is presented for chest radiographs. Potential candidates are detected using a Weighted Multiscale Convergence-Index filter. Later, candidates are segmented using adaptive distance-based threshold algorithm. In [6], Support Vector Machine (SVM), kernel Fisher discriminant and Adaboost classification methods are employed for CAD of lung nodules. In [7], a neural classifier to reduce the False Positives (FPs) is implemented. In [8], various classification techniques such as Fisher Linear Discriminant (FLD), quadratic and linear are compared. Research work presented in [9] provides the initial validation and implementation of deep learning in CAD systems for pulmonary nodule detection and diagnosis. Nodules are artificially simulated by rotations for classification using deep learning in [10] to classify them as benign or malignant. In [11], various geometric descriptors are compared with deep learning approaches for classifying nodules as benign or malignant. In [9], feature based classifiers have proven to be more effective when compared to existing deep learning techniques for CAD of lung nodules in CT scans. In [12-15], the most recent research developments based on feature-based classification for CAD systems is presented. In [15], performance analysis of CAD system at different slice thicknesses is presented for the publicly available Lung Nodule Analysis 2016 (LUNA16) dataset [16,17]. Research work presented in [15] would serve as the benchmark for our paper.

In this paper, we implement a SVM based classification approach for classification of lung nodules and compare its performance with existing benchmark. A suite of 503 features

as utilized in [4] is used in this paper. Not much research has been implemented on selection of features for classification using SVM classifier, we present a computationally efficient approach for the same in this paper. Results are presented for a publicly available LUNA16 dataset thereby setting a benchmark for future research efforts.

The remainder of this report is organized as follows. Section 2 provides a brief description about the databases that are employed for this research. Section 3 describes the CAD system architecture adopted in this paper. Section 4 elucidates the classification methods along with the feature selection algorithm for SVM classifier. Experimental results are presented in Section 5. Finally, conclusions are offered in Section 6.

II. Materials

In this research, we utilize a publicly available dataset present for the LUNA16 grand challenge [16, 17]. Subset of dataset presented for Lung Image Database Consortium – Image Database Research Initiative (LIDC- IDRI) [18] is utilized for this grand challenge. LUNA16 is comprised of 888 CT scans with 1351 radiologists' markings as nodules. Four radiologists independently studied each CT scan and marked all the suspicious markings. Annotations above 3mm marked by three of the four radiologists are considered for the challenge. For this research, we remove all the redundant markings by different radiologists for a single target nodule. In total, we have 1141 target nodule cue points.

III. CAD System Architecture

We adopt the CAD system presented in [3-5] for this research and its corresponding block diagram is provided in Figure 1. At first, lung segmentation is implemented using an active shape model [3]. Later, nodules are detected and segmented simultaneously using multiple gray level thresholding and morphological operations [3]. A set of 503 features is later computed to classify the candidate as a nodule or non-nodule [4]. Features are selected for classification using SFS method based on 10-fold validation of the training data. Area under the Free-response Receiver Operating Characteristic Curve (FROC) from 0-10 FPs is used as the performance metric. In [3-5], classification is performed using a FLD classifier. In this paper, we compare the performance of the FLD classifier presented in [5] with a SVM based classifier architecture as described in Section 4.

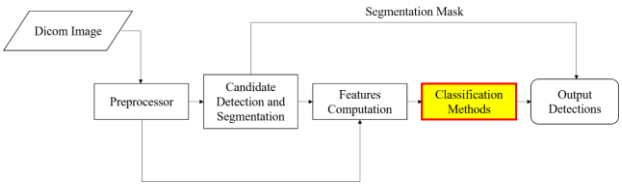

Figure 1: CAD System Block Diagram adopted from [3]

IV. Classification Methods

After the computation of 503 features for each potential candidate, we shortlist them to 300 based on rank. Ranking of features is implemented to filter a subset of features to assist the feature selection process. Features are ranked computed using MATLAB built-in function 'rankfeatures()' [20] with the 'roc' option. As implemented in [3, 4], we select a subset of features from the shortlisted ones by SFS. This method is implemented solely based on the training dataset using 10-fold validation technique and a FLD classifier.

After the selection of features using SFS method, we adopt the knee point method used in [3-5] to determine the set of features selected for classification step. Knee point has proven to be a highly effective method for classification [3-5]. Knee point in the SFS metric curve is the number of features at which the classifier achieves a good training performance but does not saturate thereby avoiding overfitting. We study the performance of both FLD classifier and SVM classifier with linear kernel using this approach. Note that, we do not perform any separate feature selection for SVM classifier due to the computational complexity associated with it. Also, FLD classifier forms a linear boundary between the two classes and a linear SVM effectively does a piecewise linear boundary. Hence, we believe SFS based on FLD can be effective for classification using SVM. Figure 2 presents the block diagram for the same.

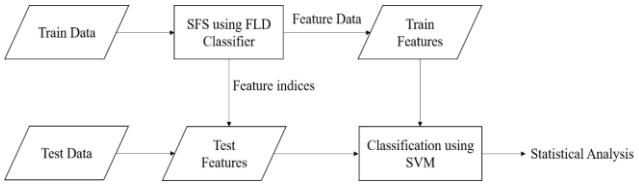

Figure 2: Block Diagram for SVM classification based on SFS using FLD classifier

In knee point strategy, typically only 10-15 features are selected for classification. So, we study the performance of both FLD and SVM classifiers by choosing a relatively larger set of features selected using SFS approach as it has been proven in the literature that SVM has the capability to form a well-defined boundary with a relatively higher suite of features. In addition, we study the SVM performance as a function of features selected using rank and later compare using our proposed feature selection approach for SVM classification.

V. Experimental Results

In this section, we present the experimental results obtained for various methods proposed in this paper. At first, we present the candidate detection results based on methods described in Section 3. Figure 3 presents a typical candidate detector result obtained for a specific slice from the 'sub2_P33' case from the LUNA16 database. Figure 3 clearly indicates that our candidate detector successfully detected the nodule marked by a radiologist. Among the 1141 target nodule cue points present in LUNA16 database, our candidate detector successfully detected 1031 cues with an accuracy of 90.35%. Table 1 presents the

candidate detection results in terms of both specificity and sensitivity.

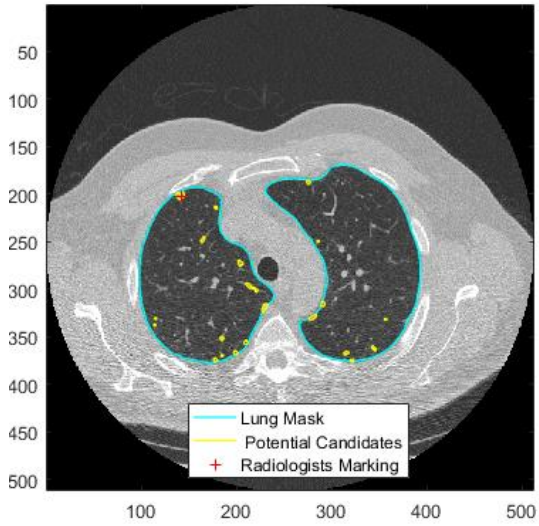

Figure 3: Typical Candidate Detector Result for the case 'sub2_P33'

Table 1: Candidate Detector Performance for LUNA16 database

Number of Cases	Number of Candidates Detected	Number of Target Nodules	Number of Nodules Detected
888	848383	1141	**1031**

LUNA16 grand challenge has divided their database into 10 different subsets (subsets 0-9). We utilize the same set of indices for our 10-fold validation results in this paper. For testing each subset, we make sure to exclude it for training purposes. Training step includes feature selection and classifier training. For instance, if we are testing subset 0, we make sure to train solely based on subsets 1-9. For the feature selection step, we utilize all the nodules detected by our candidate detector but use only 20% of non-nodules (randomly selected) to reduce time consumption. The performance of feature(s) is measured based on 10-fold validation based on AUC from 0-10 FPs in FROC using a FLD classifier for both FLD and SVM classification. A point to note, 'StandardSeparation3d' is seeded as the first feature for all SFS processes. For the first experiment, we select the knee point (subjectively determined) in the performance metric curve for each subset and compare the performances. Number of features selected using knee point strategy are in range of 10-15. We make sure to incorporate all the potential candidate detections for classification purposes. Later, we select 50 features based on SFS method of feature selection for classifying each subset. 50 features are selected due to minimal change in training performance after selecting 50 features (in the order of 10^{-4}) for all subsets. Figure 4 compares the FROC results obtained using FLD and SVM classifier for each feature set. Table 2 summarizes the results using both feature set for the two classification methods. FROC results are summarized in terms of AUC from 0 – 10 FPs. Also, results are summarized in terms of scoring metric proposed in ANODE 2009 [19] and LUNA16 [16, 17] grand challenge. This scoring metric is computed based on the average sensitivity at 7 different points: 0.125, 0.25, 0.5, 1, 2, 4 and 8 FPs from the FROC curve. We also measure the sensitivity at 3 FPs which is usually the operating point for the radiologists.

Figure 4 and Table 2 clearly indicate that linear SVM with 50 features provides the best performance amongst all the classification methods presented. Performance is closely followed by a FLD classifier with 50 features.

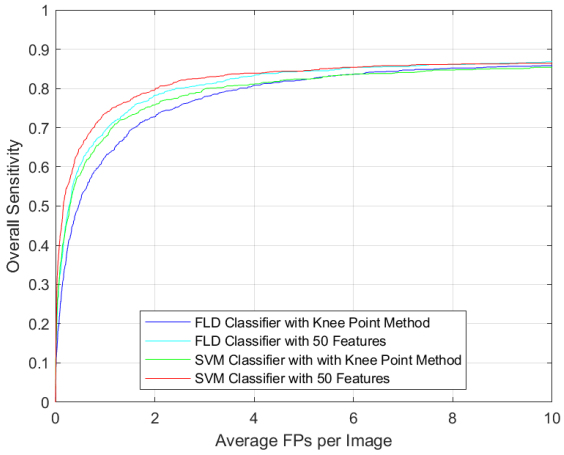

Figure 4: FROC Comparison of CAD Performance for All Classification Methods

Table 2: Comparison of CAD Performance for Various Classification Methods

Classification Method	Number of features	AUC (0- 10 FPs)	ANODE Scoring Metric	CAD Sensitivity at 3 FPs
FLD	Knee Point	7.74	0.60	77.91
SVM	Knee Point	7.88	0.64	79.84
FLD	50	8.04	0.66	80.08
SVM	**50**	**8.16**	**0.70**	**82.82**

Figure 5 and Table 3 present the results obtained using SVM classifier as a function of top rank features (50,100, 150 and 200).

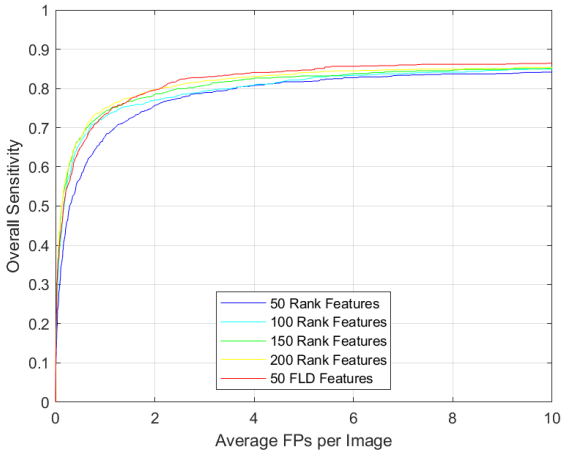

Figure 5: SVM Performance Comparison using different Feature sets

Table 3: Comparison of CAD Performance by using SVM classifier as a function of Rank features

Number of Rank Features	AUC (0- 10 FPs)	ANODE Scoring Metric	CAD Sensitivity at 3 FPs
50	7.80	0.64	78.88
100	7.95	0.69	79.32
150	8.05	0.71	80.81
200	8.11	0.71	81.95

VI. CONCLUSIONS

In this paper, we have presented a novel classification approach for CAD of lung nodules in CT scans. Detailed performance analysis is provided for a publicly available dataset thereby setting a benchmark for future research efforts. We have also presented a computationally efficient feature selection method for a linear SVM classifier. SFS based on FLD classification method helps in determining an optimal suite of features for classification using linear SVM at a much faster rate. We also studied the performance of SVM classifier as a function of rank features.

Table 2 clearly indicates that all performance metrics in this paper follow the same trend. SVM provides the best results when compared to FLD classifier in both scenarios i.e., when the number of features is selected based on knee point in the performance metric curve and when 50 features are selected for classification. Also, SVM forms a well-defined boundary using a large set of features at a much faster rate.

Figure 5 and Table 3 clearly indicate that SVM performs better with more rank features. However, SFS based on FLD for SVM classification slightly outperforms SVM classifier designed with 200 rank features.

An area of future research would be to optimize the feature set for SVM classification method by performing SFS based on SVM. However, this method would be computationally complex, memory demanding and time consuming, especially for CT scans. With the advancement of supercomputers, this could be possible. Another area of future research would be to study the performance of deep learning and featureless approaches for CAD of lung nodules.

REFERENCES

[1] The American Cancer Society, Cancer Facts and Figures 2018.

[2] Henschke, C.I., McCauley, D.I., Yankelevitz, D.F., Naidich, D.P., McGuniness, G. Miettinen, O.S., Libby, D.M., Pasmantier, M.W., Koizumi, J., Altorki, N.K., Smith, J.P., 1999. "Early cancer action project: overall design and findings from baseline screening". Lancet, 354 (1973), 99-105.

[3] Messay, T., Hardie, R. C., & Rogers, S. K., (2010). "A new computationally efficient CAD system for pulmonary nodule detection in CT imagery". Medical Image Analysis, 14(3), 390-406.

[4] Narayanan, B. N., Hardie, R. C., Kebede, T. M., & Sprague, M. J. "Optimized feature selection-based clustering approach for computer-aided detection of lung nodules in different modalities". Pattern Analysis and Applications, 1-13, 2017.

[5] Hardie, R. C., Rogers, S. K., Wilson, T., & Rogers, A. (2008). "Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs". Medical Image Analysis, 12(3), 240-258.

[6] Wei, L., Yang, Y., Nishikawa, R. M., & Jiang, Y. (2005). "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications". IEEE transactions on medical imaging, 24(3), 371-380.

[7] Gori, I., Fantacci, M. E., Martinez, A. P., & Retico, A. (2007). "An automated system for lung nodule detection in low-dose computed tomography". Medical imaging, International Society for Optics and Photonics, 65143R-65143R.

[8] Narayanan, B. N., Hardie, R. C., & Kebede, T. M. (2016). "Analysis of various classification techniques for computer aided detection system of pulmonary nodules in CT". In Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS), 88-93.

[9] Gruetzemacher, R., & Gupta, A. (2016). "Using deep learning for pulmonary nodule detection & diagnosis". Twenty-second American conference on information systems, San Diego.

[10] Hua, K. L., Hsu, C. H., Hidayati, S. C., Cheng, W. H., & Chen, Y. J. (2015). "Computer-aided classification of lung nodules on computed tomography images via deep learning technique". OncoTargets and therapy, 8.

[11] Yang, H., Yu, H., & Wang, G. (2016). “Deep Learning for the Classification of Lung Nodules”. arXiv preprint arXiv:1611.06651.

[12] Shaukat, F., Raja, G., Gooya, A., & Frangi, A. F. (2017). “Fully automatic and accurate detection of lung nodules in CT images using a hybrid feature set”. Medical Physics.

[13] Jaffar, M. A., Siddiqui, A. B., & Mushtaq, M. (2017). “Ensemble classification of pulmonary nodules using gradient intensity feature descriptor and differential evolution”. Cluster Computing, 1-15.

[14] Liu, J. K., Jiang, H. Y., He, C. G., Wang, Y., Wang, P., & Ma, H. “An Assisted Diagnosis System for Detection of Early Pulmonary Nodule in Computed Tomography Images”. Journal of medical systems, 41(2), 30, 2017.

[15] Narayanan, B. N., Hardie, R. C., & Kebede, T. M. “Performance analysis of a computer-aided detection system for lung nodules in CT at different slice thicknesses”. SPIE Journal of Medical Imaging, 5(1), 014504, doi: 10.1117/1.JMI.5.1.014504, 2018.

[16] Setio, A. A. A., Traverso, A., de Bel, T., Berens, M. S., Bogaard, C. V. D., Cerello, P., & van der Gugten, R. “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge”. Medical Image Analysis, 42, 1-13, 2017.

[17] Lung Nodule Analysis, 2016. https://luna16.grand-challenge.org/home. Accessed April 3, 2018.

[18] Armato III, S. G., McLennan, G., McNitt-Gray, M. F., Meyer, C. R., Yankelevitz, D., Aberle, D. R., & Reeves, A. P. “Lung image database consortium: Developing a resource for the medical imaging research community”. Radiology, 232(3), 739-748, 2004.

[19] van Ginneken, B., Armato, S. G., de Hoop, B., van Amelsvoort-van de Vorst, S., Duindam, T., Niemeijer, M., & Camarlinghi, N. “Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study”. Medical image analysis, 14(6), 707-722, 2010.

[20]https://www.mathworks.com/help/bioinfo/ref/rankfeatures.html. Accessed April 3, 2018.