Prediction of Lung Nodule Characteristic Rating using Best Classifier Model

Vinay K
DoS in Computer Science
Manasagangothri
University of Mysore
Mysore, INDIA

Ashok Rao
Freelance Academician
165, 11th main
Saraswathipuram
Mysore, INDIA

G. Hemantha Kumar
DoS in Computer Science
Manasagangothri
University of Mysore
Mysore, INDIA

ABSTRACT
In this paper, we are exploring the response of individual classifier families on imbalanced medical data. In this work we are using LIDC (Lung Image Database Consortium) dataset, which is a very good example for imbalanced data. The main objective of this work is to examine how will be the response of different categories of classifier on imbalanced dataset. We are considering five categories of dataset which are grouped as, Instance Based classifier, Rule Based classifiers, Functional Classifier, Decision Tree classifier and Ensemble of Classifiers. The results from our experiments will be evaluated based on following performance metrics such as Accuracy, Precision, Recall, F-measure, Area under curve and kappa statistics.

Keywords
Ensemble of classifier, Decision Tree, Kappa Statistics

1. INTRODUCTION
Based on the GLOBOCAN 2008 estimates, about 13.7 million cancer cases and 9.6 million cancer deaths are estimated to have occurred in 2008 of these, 56% of the cases and 64% of the deaths occurred in the economically developing world [1]. Lung cancer is the most frequently diagnosed cancer and the leading cause of cancer death. Survival from the lung cancer is directly related to early and correct detection and diagnosis of the malignant lesions. Studies show that positive diagnosis from radiologists is possible to the maximum accuracy of 70% – 80% when it has been diagnosed using computerized tomography (CT) imaging. Hence usage of CT screening technique is widely used across the world. The possibility of survival rate from cancer is very less and mortality rates are increasing year to year. The main failure for such mortality is due to wrong diagnosis of cancer disease. The early detection of cancerous nodule will surely helps in curing the disease. It is human tendency to make error in manually diagnosing the lesions as nodule or nodule. Many cases have different interpretation between the radiologists. This is generally true of biomedical field, where opinions are extremely subjective. Studies have shown that radiologist frequently fail to agree with all nodules, especially in marginal cases and the examination of CT scan is time consuming and error prone task and its human tendency to make mistakes due to large work pressure [2]. The main purpose of Computer Aided Diagnosis (CAD) systems is to assist radiologist in medical decision making.

2. RELATED WORK
Ekrain et.al [3] investigated several approaches to combine delineated boundaries and ratings from multiple observer and they have used p-map analysis with union, intersection and threshold probability to combine the boundary reading and claimed that threshold probability approach provides good level of agreement. Lee et al [4] proposed a method using two step approaches for feature selection and classifier ensemble construction. They used genetic algorithm in initial round of feature reduction and claimed that use of ensemble classifiers that explicitly enable classification using multiple different subsets helps to relief the problem of selection of high performance feature subset in developing CAD system. Various classifier models have been used for lung nodule classification. Linear classifiers are popular due to their speed and accuracy, including Artificial Neural Network (ANN) [5]. Lee et al [6] have developed a CADx system based on two-step feature selection and advanced classifier algorithm.

Nakumura et al [7] worked on simulating the radiologists perception of diagnostic characteristic rating such as shape, margin, irregularity, Spiculation, Lobulation, texture etc. on a scale of 1 to 6 and they extracted various statistical and geometric image features including fourier and radiant gradient indices and correlated these features with the radiologists ratings. They showed correlation between radial gradient indices with spiculation and the other geometric features with shape and concluded that there was poor predictive performance in ratings of radiologists due to variability in inter observer ratings. Ebadollahi et al [8] proposed a framework that uses semantic methods to describe visual abnormalities and exchange knowledge with medical domain.

3. METHODS AND MATERIALS
The dataset used in this work is obtained from LIDC (Lung image database Consortium) [9]. LIDC is lung image dataset which is publically available through National cancer Institute's Imaging Archive. In this section we briefly discuss about the procedure of data collection process by LIDC. The dataset comes with CT images and XML file. CT images are of DICOM (Digital Imaging and Communications in Medicine) image format and it is a standard for handling, storing, printing, and transmitting information in medical imaging. XML file gives us the information about the nodule details such as, its nodules spatial location co-ordinates, size and radiological characteristic ratings. LIDC have four radiologists in the panel. LIDC data collection will be made of two phases namely Blinded session and unblinded sessions of reading. In the first phase each radiologist reviews CT scan independently. In the second phase ratings from all the four radiologists are gathered together and presented to each other radiologists for a second review, allowing radiologist to refine their earlier opinion based on others review. Subsequently results of each radiologists were compiled later to form final unblinded review.

As mentioned before LIDC data comes along with XML file which hold the information about the nodules for the communication of results. XML files holds the spatial location information about the three types of lesions, they are nodules < 3 mm, nodules > 3 mm and Non-nodules > 3mm of diameter as marked by panel of four radiologists. Apart from spatial location information, XML files also gives information about the radiologist's ratings for nine nodule characteristics: Lobulation, internal structure, calcification, subtlety, spiculation, margin, sphericity, texture and malignancy.

In this work 124 out of 399 cases from LIDC dataset has been considered. We have extracted 4532 nodule from those 124 cases. The samples of nodules which we have extracted from the CT images are shown in figure below (see Fig 1).

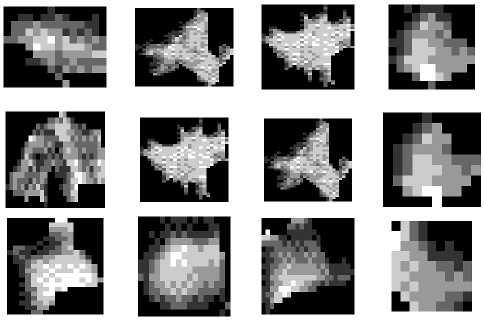

Fig 1: Samples of Nodules Extracted

4 FEATURE EXTRACTION

Feature extraction plays a very important role in classification. We have extracted numerous features that can be grouped as low level image features wiz, Size features, shape features, intensity features and texture features.

For size features we have considered Area, Convex Area, Perimeter, Convex Perimeter, EquivDiameter, MajorAxisLength and MinorAxisLength.

For shape features we have considered, Circularity, Roughness, Elongation, Compactness, Eccentricity, Solidity and Extent.

The four intensity features which we have extracted from lung nodule images are Minimum Intensity, Maximum Intensity, Mean Intensity and Standard Deviation Intensity.

For texture features have been further grouped into two categories based on the approach used for extraction of features. They are statistical based texture features and Transform based texture features. Statistical methods describe the image using pure numerical analysis based on pixel intensity values, where as Transform based approaches perform transformation to the original image by filtering and obtaining the response image, which is later analyzed as a representative for the original image.

For statistical based approach we have used haralick features. A haralick feature is a group of 13 image features which were extracted as follows. Co-occurrence matrix of the input image is calculated along four directions ((0°, 45°, 90° and 145°) and five distances (1, 2, 3, 4 and 5). This yields 20 response matrices on which haralick features are extracted. The thirteen haralick features which were extracted in this work are Energy, Correlation, Inertia, Entropy, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Difference, Average, Difference Variance, Difference Entropy, Information measure of correlation 1 and Information measure of correlation 2

Gabor feature extraction approach is also being used in this work for extraction of texture features under category of transform based features. Gabor features is a well known method which extracts texture information from an image in the form of response image [10]. We calculated this at four orientations (0°, 45°, 90° and 145°) and three frequencies (0.3, 0.4 and 0.5) by convolving the image with 12 Gabor filters. Here we have considered only the resulting mean and standard deviation of 12 Gabor response images thus resulting in 24 Gabor features per nodule image.

The detailed low level images features which have been considered in this work are listed in Table 1.

Table 1: Low level image features

Size Feature	Shape Feature	Intensity Feature
Area	Circularity	MinIntensity
Convex Area	Roughness	MaxIntencity
Perimeter	Elongation	MeanIntensity
Convex Perimeter	Compactness	SDIntensty
EquivDiameter	Eccentricity	
MajorAxisLength	Solidity	
MinorAxisLength	Extent	
Texture Features		
24 Gabor features are mean and standard deviation of 12 different gabor response images at orientation = 0, 45, 90, 135 and time frequency = 0.3, 0.4,05 13 Haralick features calculated from co-occurrence matrices. Energy, Correlation, Inertia, Entropy, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Difference, Average, Difference Variance, Difference Entropy, Information measure of correlation 1, Information measure of correlation 2		

In this work we have grouped image features into two categories (1) low level image features (2) radiologist's characteristic ratings provided by LIDC. As we explained in the earlier section we have extracted fifty five low level features which are concatenated with eight radiological predictions. Therefore the total number of features we are considering in this work equals to sixty three.

The overview of the dataset and number of features considered in this work is given in Table 2.

Table 2: Overview of dataset considered in this work

Dataset and Feature Extraction Details	
No. of cases considered	124
No. of Instances	14956
No. of Nodules	4532
No. of low level image features extracted	55
No. of radiologist characteristic ratings considered	08
Total No. of features used in the work	63

5 EXPERIMENTAL SETUP

In this work we have carried out different set of experiments on the same dataset with different parameter setup to observe the performance of the classifier on imbalanced dataset. As we mentioned before in this work we are using five different classifier families to carry out our experiments. We used K-Nearest Neighbor (KNN) under instance based classifier, Multi Layer Perceptron (MLP) and Support Vector Machine (SVM) under functional based classifiers, PART and RIDOR for rule based classifiers, J48 and REPTree under decision tree classifiers. We are considering Bagging and Boosting under Ensemble Methods for our experiments.

Top performing classifier in the list of instance based, functional based, rule based and decision tree based classifiers are chosen to be a base classifiers for Bagging and AdaBoost methods.

6 RESULTS AND DISCUSSIONS

The performances of classifiers are evaluated using five performance metrics which are categorized into three groups. Accuracy and F-measure are considered under threshold metrics and we have fixed threshold to 0. The classifier which performs above the threshold (> 0.5) is considered to be good performer and classifier whose performance below is threshold (< 0.5) is regarded as under performer. RMSE (Root Mean squared Error) is used as probability metric. Probability metric are minimized when the predicted value for each case is equals to true conditional probability.

Lower the RMSE value will be better the performer. AUC (Area Under Curve) is used as a rank metric and this metric measures how well the positive cases and negative cases are ordered and viewed. Kappa statics is used as agreement measures, which in turn reflect how well model agrees between the expert prediction and machine prediction. The kappa interpretation scale has been given in Table 3.

Table – 3: Kappa Statistics interpretation scale

K - value	Strength of Agreement
<0	Poor
0 – 0.2	Slight
0.21 – 0.4	Fair
0.41 – 0.6	Moderate
0.61 – 0.8	Substantial
0.81 - 1	Almost perfect

The best performing classifiers are highlighted in the results table (refer Table 3) (bold cases in the column under specified performance metric). We have observed from our experimentation that the ensembles of classifiers are performing well in all the cases. In our case Bagging with REPTree base classifier has been provided excellent results when compared to all other classifier group. When we compare the Bagging with near competitor i.e., J48 decision tree performance, the results look similar. This is because the way we choose base classifier. We have chosen base classifier in such a way that it should be top performer in the list.

7 CONCLUSION

In this work we have addressed the role of classifier on medical data which happens to be imbalanced in many cases. The results from our experiments showed that ensemble of classifier approach will give good results when compared to other family of classifiers. It is worth notice that though SVM is regarded as good classifier in the pattern recognition literature is worst performer in our case. Hence it is very sensible issue in choosing classifier while dealing with imbalanced dataset. The fact behind the better performance from ensemble of classifier family is the way they classify the test examples is very much similar to assessing the label from different experts. That is ensemble of classifiers works on combination rules such voting which refers to winner take all policy. As in medical domain there is always requirement for getting many opinions and finally concluding the result based on outputs of first level. Hence the ensemble of classifier model is best suited for the fields like medical where often it is required to deal with imbalance dataset.

REFERENCES

[1] Global cancer statistics, Ahmedin Jemal, Freddie Bray, Melissa M. Center, Jacque Ferlay, Elizabeth Ward, David Forman, A cancer Journal for clinicians. (2008).

[2] Miles N.Wernick, Yongyi Yang, Jovan G. Brankov,

[3] Ekrain Varutbangkul, Vesna Mitrovic, Daniel raichu, Jacob Furst. Combining Boundaries abd Rating from Multiple Observers for Predicting Lung Nodule Characteristics. In: IEEE International Conference on Biocomputing, Bioinformatics and Biomedical technologies, 82-87 (2008).

[4] Michael C. Lee, Lilla Boroczky, Kivilcim Sungur Stasik, Aaron D.Cann, Alain C. Borczuk, Steve M. Kawut, Charles A. Powell.: Computer-aided diagnosis of Pulmonary nodules using two-step approach for feature selection and classifier ensemble construction In Artificial Intelligence in Medicine, Elsevier 50 (2010) 43-53.

[5] Katsumi Nakamura, Hiroyuki Yoshida,Roger Engelmann, Heber MacMahon,Shigehiko Katsuragawa, Takayuki Ishida, Kazuto Ashizawa and Kunio Doi, "Computerized Analysis of the Likelihood of Malignancy in Solitary Pulmonary Nodules with Use of Artificial Neural Networks", Radiology 2000; 214:823-830.

[6] Lee, M.C.; Boroczky, L.; Sungur-Stasik, K.; Cann, A.D.; Borczuk, A.C.; Kawut, S.M.;Powell, C.A.;Philips Res. North America, Briarcliff Manor, NY "A Two-Step Approach for Feature Selection and Classifier Ensemble Construction in Computer-Aided Diagnosis ", 21st IEEE International Symposium on Computer-Based Medical Systems, 2008.

[7] Nakumura K, yoshida H, Engelmann R. MacMahon H, Kasturagawa S. Ishida T, et al. computerized analysis of the likliihood of malignancy in solitary pulmonary nodules with use of artifial neural networks. Radiology 823-30. (2000)

[8] Ebadollahi, S., Johnson, D.E.., Diao M, Retrieving clinical cases through a concept space representation of text and images. SPIE Med. Imaging Symp (2008)

[9] https://ncia.nci.nih.gov (Lung Image Database Consortium)

[10] Dmitriy Zinovev, Daniela Raicu, Jacob Furst and Samuel G. Armato, "Predicting Radiological Panel Opinions Using Panel of Machine Learning Classifiers", Algorithms 2009, 2, 1473-1502; doi:10.3390/a2041473.

[11] Vinay. K, Ashok Rao, Hemantha Kumar. G "Comparative Study on Performance of Single Classifier with Ensemble of Classifiers in Predicting Radiological Experts Ratings on Lung Nodules", Indian International Conference on Artificial Intelligence (IICAI-11). ISBN: 978-0-9727412-8-6, pp 393 – 403

[12] Vinay. K, Ashok Rao, Hemantha Kumar. G, "Computerized Analysis of Classification of Lung Nodules and Comparison between Homogeneous and Heterogeneous Ensemble of Classifier Model", 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, 978-0-7695-4599-8/11, IEEE DOI 10.1109/NCVPRIPG.2011.56, pp 231-234. .

Table 4: Results from experiments

Classifier Type	**Classifier**	**Threshold Metrics**		**Probabilistic Metric**	**Rank Metric**	**Agreement Metric**
		Accuracy	**F-measure**	**RMSE**	**AUC**	**Kappa**
Instance Based	**KNN**	66.88	0.77	0.30	0.93	0.58
Function Based	**MLP**	69.21	0.81	0.32	0.91	0.60
	SVM	58.65	0.75	0.36	0.90	0.46
Rule Based	**PART**	76.98	0.84	0.29	0.92	0.71
	RIDOR	60.66	0.78	0.40	0.86	0.49
Decision Tree Based	**J48**	81.01	0.86	0.26	0.94	0.76
	REPTree	69.84	0.80	0.30	0.95	0.61
Ensemble Based	**Bagging**	**83.29**	**0.89**	**0.22**	**0.98**	**0.79**
	AdaBoost	74.08	0.83	0.27	0.96	0.67