Classifiers in Context: Prediction of Radiological Characteristic Ratings for Lung Nodule Malignancy

Vinay. K
DOS in Computer Science
Manasagangotri
University of Mysore
Mysore, INDIA

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

G. Hemantha Kumar
DOS in Computer Science
Manasagangotri
University of Mysore
Mysore, INDIA

ABSTRACT

In this paper, we are exploring a panel of classifier response to an imbalanced medical data set. 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 the response of different categories of classifier is, when subjected to imbalanced dataset. We are considering five categories of classifiers 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 performance metrics such as Accuracy, Precision, Recall, F-measure, Area under curve and Kappa statistics.

Keywords

Ensemble of classifiers, Decision Tree, Kappa Statistics

1. INTRODUCTION

Based on the GLOBOCAN 2008 estimates, about 13.7 million cancer cases and 9.6 million cancer deaths 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 also the leading cause of cancer death. This is particularly more so in economically developing world. 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 help in curing the disease in larger percentage of cases. It is natural human tendency to make error in manually diagnosing the lesions as nodule or non-nodule.

Many cases have different interpretation between the radiologists. This is generally true of much of diagnosis in biomedical field, where opinions are generally subjective even ranging in extremes. 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, more so in marginal cases, where decision making is more difficult.

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 have used genetic algorithm in initial round of feature reduction. From the obtained results they have claimed that use of ensemble of classifiers that explicitly enable classification using multiple different subsets 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

Lung Image Database Consortium (LIDC) [9] provides lung CT image data which is publically available through National cancer Institute’s Imaging Archive (web site – http://ncia.nci.nih.gov). This dataset consists of image data, radiologist’s nodule outline details and radiologist subjective characteristic ratings. The LIDC dataset currently contains complete thoracic CT scans of 399 patients acquired over different periods of time. LIDC data download comes with DICOM image and the nodule information in the XML file. This has information regarding the spatial location information about three types of lesions; they are nodules < 3 mm; nodules > 3 mm and non-nodules > 3 mm in maximum diameter as marked by panel of 4 expert radiologists. For any lesion greater than 3 mm in diameter, XML file contains spatial coordinates of nodule's outline. Since the number of radiologist in LIDC panel is 4, it is obvious that each nodule > 3mm has 4 nodule outlines. Moreover, any radiologist who identifies a nodule > 3mm also provides subjective ratings for 9 nodule characteristics, wiz.,: Lobulation, internal structure, calcification, subtlety, spiculation, margin, sphericity, texture and malignancy.

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

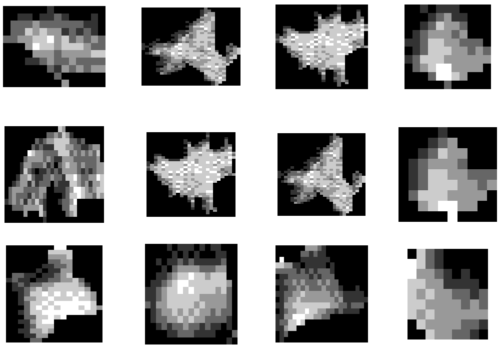

Fig 1: Samples of Nodules Extracted

4. ROLE OF IMBALANCED DATASET

Classifier performance is directly proportional to the balance in dataset [13]. David Cieslak and Nitesh Chawla [14] have mentioned in their work that many machine learning applications in areas like finance, medicine, and risk management suffer from class imbalance. The imbalance in the dataset further create complication in training the model. This in turn causes testing samples to differ significantly in their respective class distributions from those of training samples. Eventually this leads to poor classification in general and if it is good then it is likely to be biased classification.

There are two methods to deal with such situations. One is of removing the imbalances in data set and performing classification. The other is of trying to get a more objective classification by deploying panel of classifiers. The latter method is what we follow here.

As mentioned before LIDC dataset is good example for such imbalance, which we can see in the table 1.

Table – 1 Malignancy sample distributions in dataset

Class Label	No. of Samples
Highly Unlikely	572
Moderately Unlikely	733
Indeterminate	1285
Moderate	796
Suspicious	1146

Our work is mainly focused on predicting rating for this malignancy. The above table gives the instance distribution for malignancy case. The rating for the malignancy is further divided into multiclass such as Highly Unlikely, Moderately Unlikely, Indeterminate, Moderate and Suspicious cases. As we can see in table that the number of samples for highly likely cases is 572 where as for the cases Moderately Unlikely and Suspicious are 1285 and 1146 respectively. It means the number of cases for Moderately Unlikely and Suspicious is almost the double the number of samples in Highly Unlikely cases. In such scenario when we classify such imbalanced dataset, though using good performing classifier will result in bias. Since the classifier will get more number of samples of some classes and it will get fewer number of samples of other classes. Hence the classifier tends to get biased towards the case which has more number of samples. It is to be noted that majority of real life medical data is indeed imbalanced. This reflects the distribution of such issues across the general population. Thus working on such data is important since it captures realistic situation much more effectively.

5. FEATURE EXTRACTION

In this proposed work we consider same set of features which we have calculated in our earlier work [11] , [12]. Our feature set consists fifty five two dimensional, low level image features grouped into four categories: size feature, shape feature, intensity feature and texture features. Further we consider the nodule which has largest area and the image feature are extracted only for this largest nodule. Table 1 gives the details about low level image features that we have used in this work.

As size features, we are using nodule area, equivdiameter, axis length etc,. Main shape features are circularity, solidity, eccentricity and elongation etc., and Intensity features are minimum intensity, maximum intensity, their mean intensity and their standard deviation intensity levels. As texture features, we are extracting feature from nodule region using two approaches: a statistics based method and a transform based method. Statistical methods describe the image using pure numerical analysis of pixel intensity values. Transform based approaches perform transformation to the original image by filtering and obtaining response image, which is later analyzed as a representative for the original image. Here we have used Haralick features (a statistics based method) and Gabor filters (a transform based method).

Co-occurrence matrix were calculated along four directions (0°, 45°, 90° and 145°) and five distances (1, 2, 3 ,4 and 5). Once the co-occurrence matrices are calculated, thirteen Haralick features are calculated from each matrix and we averaged the features along all direction and distances resulting in 13 haralick descriptors per nodule image.

Gabor filtering is a well known transform based methods which extracts texture information from an image in the form of response image. 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 mean and standard deviation of 12 Gabor response images thus resulting in 24 features per nodule image [10].

The detailed low level images features which have been considered in this work are listed in Table 2.

Table 2: Low level image features

Size Feature	Shape Feature	Intensity Feature
Area Convex Area Perimeter Convex Perimeter EquivDiameter MajorAxisLength MinorAxisLength	Circularity Roughness Elongation Compactness Eccentricity Solidity Extent	MinIntensity MaxIntencity MeanIntensity SDIntensty
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		

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 3.

Table 3: 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

6. EXPERIMENTS 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.

7. RESULTS AND DISCUSSION

The performances of classifiers are evaluated using five performance metrics which are categorized into four groups. These are threshold metric, probabilistic metric, rank metric and agreement metric. 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, 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 4.

Table – 4: 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 4) (bold cases in the column under specified performance metric). We have observed from our experiments that the ensembles of classifiers are performing well in all the cases. In our case Bagging with REPTree base classifier has 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. The main observation we have noticed from the first set of experiments is that decision tree classifiers such as J48 and REPTree classifiers perform better when compared to other single classifier from other classifier family.

In the other set of experiments we have investigated the performance of the ensemble of classifiers on our dataset. The ensemble methods viz., Bagging and AdaBoost were used in experiment and base classifiers for these methods are J48 and REPTree, which are the top performing classifiers from our previous experiment. A good ensemble of classifiers can be constructed when there is much diversity provided to the ensemble. This diversity can be achieved using the base classifiers which are unstable in nature and prone to error with minimum changes in the input level.

The choice of using decision tree classifier as a base learner is because of their instability in nature to small changes in input parameter. In literature it has been claimed that decision trees are unstable classifiers [ ]

We have also cross checked the methods with different combination of base classifiers. The results from experiment (see table 7) shows that ensemble of classifiers performs better compare to single model. But the above statement contradicts when we talk about SVM classifier. Since SVM performs very well on balanced data. On the other hand it performs very poorly either in case of single classifier model or ensemble of SVM classifier model. Our results from experiments using SVM classifier show that SVM is a underperforming classifier for imbalanced or imbalanced dataset.

Following are the observation one can make from the results obtained:

1. **Ensemble of classifier model with decision tree as a base learner**

   - When J48 classifier (in single classifier model) used, ACC= 81.01%. When J48 classifier used as a base classifier in Bagging methods ACC = 85.14%. It shows there is a 4.13% supplementary increment of accuracy value which shows there is 4.8% improvement in the classification result.

   - In terms of other performance metrics (values obtained from experiments is given in parenthesis) such as F-measure (0.86 to 0.89), RMSE (0.28 to 0.22), AUC (0.94 to 0.98) and Kappa statistics (0.76 to 0.79) is also yielded supplementary results when J48 classifier is used in ensemble rather than in single classifier model.

   - Similar interpretation can also be done for the REPTree decision tree classifier. When it has been used in ensemble it performs much better compared to REPTree used in single classifier model. *(See table (4 and 5).*

   - Not only in case of Bagging but also in case AdaBoost method we have obtained supplementary improved results in classifier performance.

2. **Ensemble of classifier model with Rule based classifier as base learner**

   - In this set of experiment we tried to examine the performance of Ensemble method with base classifier which is other than decision tree such as PART, KNN and SVM.

   - Results with PART (bagging): ACC (76.98 to 82.66), F- measure (0.84 to 0.89), RMSE (0.29 to 0.23), AUC (0.92 to 0.98) and Kappa (0.71 to 0.78). **With AdaBoost method:** ACC (76.98 to 84.02), F-measure (0.84 to 0.90), RMSE (0.29 to 0.25), AUC (0.92 to 0.98) and Kappa (0.71 to 0.79).

   - It can be seen from the results above that PART classifier performed better when it has been used in ensemble mode, both in, Bagging and AdaBoost methods.

   - The fact behind PART classifier's good performance is also because of its instability in classification nature when there is some change in input.

3. **Ensemble of classifier model with Function based and instance based classifiers as base learner**

   - Results with KNN (Bagging): ACC (66.88 to 67.81), F- measure (0.77 to 0.79), RMSE (0.30 to 0.29), AUC (0.93 to 0.94) and Kappa (0.58 to 0.59). With KNN (AdaBoost): ACC (66.88 to 66.78), F-measure (0.77 to 0.77), RMSE (0.30 to 0.32), AUC (0.93 to 0.90) and Kappa (0.58 to 0.57).

   - KNN classifier has not given significant improvement in the results when it has been used in bagging, though it has given little improvement with respect Accuracy and F-measure. Other than this it has underperformed compare to using KNN alone. When we compare the same with AdaBoost it has performed very poor.

   - Since KNN is regarded as stable classifier, hence there is no remarkable improvements yielded when it is used as a base learner in ensemble mode.

   - Results with SVM (Bagging): ACC (58.65 to 28.21), F- measure (0.75 to 0), RMSE (0.33 to 0.53), AUC (0.90 to 0.50) and Kappa (0.46 to 0). With KNN (AdaBoost): ACC (66.88 to 28.21), F-measure (0.77 to 0), RMSE (0.30 to 0.51), AUC (0.93 to 0.51) and Kappa (0.58 to 0).

   - From results we can notice that SVM is performing poorly when it is used in single model and also it performed worst when it has been used in ensemble in either of bagging or AdaBoost case.

   - In case of SVM in ensemble method all the performance metric shows that SVM is better used in single classifier method rather than in ensemble. The Kappa statics value for SVM ensemble is 0 and which is interpreted as poor level of agreement.

   - SVM had performed badly because it is a stable classifier working very well under balanced data input. Now that it is subjected to imbalanced so SVM has failed to perform well.

8. CONCLUSION

The main focus of our work is to address the role of classifier on medical data which happens to be imbalanced as is the case frequently. The results from our experiments show that ensemble of classifier approach will give much improved results when compared to other family of single classifiers. It is worth noticing that, though SVM is regarded as good classifier in the pattern recognition literature, it is the worst performing one in our case. Hence it is a very critical 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 as voting which refers to winner take all policy. As in medical domain there is always requirement for getting multiple opinions and finally conclude based on outputs of first level. Hence the ensemble of classifier model is better suited for obtaining results in domains like medical data analysis where often it is required to deal with imbalance dataset. This is also done in real practice by way of second/multiple opinions and tests done on patients in critical cases.

9. 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, Grigori Yougnaov, Stephen C. Strother.: Machine Learning in Medical Imaging. In: IEEE signal processing Magazine (2010).

[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] National Cancer Archive website: http://ncia.nci.nih.gov

[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.

[13] Timotius, I.K, "Arithmetic means of accuracies: A classifier performance measurement for imbalanced data set", 2010, International Conference onAudio Language and Image Processing (ICALIP), 978-1-4244-5856-1

[14] David Cieslak and Nitesh Chawla, "Analyzing PETs on Imbalanced Datasets When Training and Testing Class Distributions Differ" PAKDD'08 Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining, Pages 519-526, ISBN:3-540-68124-8 978-3-540-68124-3

Table 5: Results from experiments using single classifier

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

Table 6: Results from experiments using Ensemble of Classifier

Ensemble of Classifier Model	**Base Classifier Used**	**Threshold Metrics**		**Probabilistic Metric**	**Rank Metric**	**Agreement Metric**
		Accuracy	**F-measure**	**RMSE**	**AUC**	**Kappa**
Bagging	**REPTree**	83.29	0.89	0.22	**0.98**	0.79
	J48	85.14	**0.90**	**0.21**	**0.98**	**0.81**
AdaBoost	**REPTree**	74.08	0.83	0.27	0.96	0.67
	J48	**85.47**	**0.90**	0.23	**0.98**	**0.81**

Table 7: Results from experiments using Ensemble of Classifier

Ensemble of Classifier Model	**Base Classifier Used**	**Threshold Metrics**		**Probabilistic Metric**	**Rank Metric**	**Agreement Metric**
		Accuracy	**F-measure**	**RMSE**	**AUC**	**Kappa**
Bagging	**PART**	82.66	0.89	0.23	0.98	0.78
	KNN	67.81	0.79	0.29	0.94	0.59
	SVM	28.21	0	0.53	0.50	0
AdaBoost	**PART**	84.02	0.90	0.25	0.98	0.79
	KNN	66.78	0.77	0.32	0.90	0.57
	SVM	28.21	0	0.51	0.51	0