Area under the Distance Threshold Curve as an Evaluation Measure for Probabilistic Classifiers

Sydney Williams[1], Michael Harris[2], Jacob Furst[3], and Daniela Raicu[3]

[1] Illinois Institute of Technology, Biomedical Engineering, Chicago, Illinois
sydney@hawk.iit.edu
[2] Sonoma State University, Computer Science, Rohnert Park, California
harrismi@seawolf.sonoma.edu
[3] DePaul University, Computing and Digital Media, Chicago, Illinois
{jfurst,draicu}@cdm.depaul.edu

Abstract. Evaluation for probabilistic multiclass systems has predominately been done by converting data into binary classes. While effective in quantifying the classifier performance, binary evaluation causes a loss in ability to distinguish between individual classes. We report that the evaluation of multiclass probabilistic classifiers can be quantified by using the area under the distance threshold curve for multiple distance metrics. We construct our classifiers for evaluation with data from the National Cancer Institute (NCI) Lung Image Database Consortium (LIDC) for the semantic characteristic of malignancy. We conclude that the area under the distance threshold curve can provide a measure of the classifier performance when the classifier has more than two classes and probabilistic predictions.

Keywords: Machine learning, medical informatics, probabilistic classifier, ROC curve, K-Nearest neighbor.

1 Introduction

In the advent of classifiers as key diagnostic tools within medical imaging, rigorous exploration of evaluation measures has begun in order to assess their performance. In particular, accuracy and the receiver operator characteristic curve (ROC) have been well-accepted within the machine learning community [1]. In more recent analyses, cost curves have also been reviewed because of the ability to assess different severities of misclassifications [2]. Nevertheless, previous discussions have been limited to binary labeling systems, where classifiers identify the probability of either a positive or negative label only.

While a binary label is often the preferred output of a computer aided diagnosis system (e.g., telling a doctor "healthy" or "sick"), sometimes an output that has more details is required. Multiple label classifiers can produce an output that gives a range of information to a physician, such as a rating on a 1-5 scale, or the structure of the tissue sample as a type (fat, fluid, soft tissue, air, etc). In contrast to ordinal regression where the relationship between labels 1-5 is significant, multiclass classifiers focus on class distinction, regardless of the label given. These types of classifiers are

P. Perner (Ed.): MLDM 2013, LNAI 7988, pp. 644–657, 2013.

also useful outside the medical field for purposes such as image labeling, identifying genre's in music, recommender systems, and any other problem where the desired output is to identify which of several types the input belongs to.

By using a multiclass probabilistic input, a classifier will learn using a range of information. For example, in this paper we see that multiple radiologists annotate characteristics lung nodules uniquely, and a multiclass input accounts for these differences in training. Furthermore, by combining probability with a multiclass output, one can assess how similar an object is to more than one class. For example, in showing the probabilities for the most likely cases a doctor would have more information than if simply provided a single class, further increasing the data available when compared to binary classification. Because classifiers are not perfect, having an effective way to show the most likely possibilities for the object in question is useful: a medical expert may agree with the classifiers second-most probable choice instead of the first choice.

In order to evaluate a probabilistic and multi-class classifier, a new evaluation measure has been introduced by Zinovev, Furst, and Raicu in 2011, the Area Under the Distance Threshold Curve (AUCdt) using relative differences measured by the Jeffrey Divergence [3]. Expanding upon the various evaluation functions described by Amor, Benferhat, and Elouedi in 2006 [4] the distance threshold curve helps visualize and quantify multiclass probabilistic performance. In this paper, we will expand upon the previously used AUCdt by incorporating additional distance measures as well as attempting to weight the difference across multiple class labels in regards to their relative "costs" for our application. The evaluation difference metrics that we are investigating are the City Block Difference (CB), the Jeffrey Divergence (JD), and Earth Movers Distance (EMD). Our goal is to show that these new quantities will help evaluate multiclass probabilistic classifiers just as has been done previously in the binary class case with area under the ROC curve (AUC).

2 Background

There are several challenges in evaluating probabilistic multiclass classifiers. One of these is that there is not always a correct answer, as a sample may belong to multiple classes, or there is disagreement between experts as to which class the sample belongs to. Many of the evaluation methods that are used for binary classifiers simply do not work very well when applied to multi-class output, or do not provide an accurate picture of how close the predicted output is to the expected output.

Accuracy is simply defined as the correct number of classifications over the total number of instances, or also 1-ERROR [1]:

$$Accuracy = \frac{TP+TN}{CP+CN} \tag{1}$$

where TP and TN indicate the true (correctly-classified) positives and negatives, and CP and CN , the total number of positives and negatives. While this measure is readily-available with most classifiers, its simplistic nature cannot distinguish between false positives or negatives nor generate a visual graph to define classifier performance over varying conditions, such as a threshold. Consequentially, it has been suggested that the ROC curve will allow for a more percipient yet still-consistent evaluation [5].

Discussed by Provost and Fawcett in 1997 for machine learning, the ROC curve has been a major evaluation tool for classifiers [6]. The ROC curve is constructed by plotting sensitivity vs. 1-specificity as defined using terminology from the confusion matrix seen in Table 1:

Table 1. Confusion matrix for constructing ROC curve, where labeled values come from known labels and predicted values from the classifier

Predicted \ Labeled	+	−
+	True Positive (TP)	False Positive (FP
−	False Negative (FN)	True Negative (TN)

With sensitivity being equivalent to the true positive rate:

$$Sensitivity = TPR = \frac{TP}{TP + FN} \tag{2}$$

and specificity, 1 minus the false positive rate [7]:

$$Specificity = 1 - FPR = \frac{TN}{TN+FP}. \tag{3}$$

A range of false positive probability thresholds is then created from 0 to 1. For each new threshold, if the probability of a positive label is equal or above the threshold, it is added to the value of the true positive rate and plotted as the ROC curve. Figure 1 shows an example of an ROC curve generated using artificial data with varying percentages of perfect agreement and random labels. Along with the benefit of providing an important visualization to a classifier's performance, the ROC curve also can easily be compared to "random" classifications (diagonal line with slope of 1 through curve) and is not susceptible to the bias of skewed distributions as is accuracy [8].

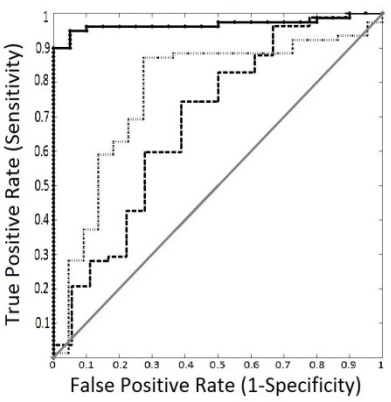

Fig. 1. Example of ROC curve— solid: artificial data with 90% agreement between predicted and labeled classes, dotted: data with 50% agreement, dashed: data with 10% agreement, diagonal: completely random performance (no classification)

Due to the benefits noted by the ROC curve, the Area Under the ROC Curve was explored by Bradley in 1997. It is calculated by taking the area enclosed beneath the ROC plot which has a range of 0 to 1. An area of 0 indicates no classification, 0.5 random classification, and 1, perfect classification [1]. The more similarity or agreement between predicted and labeled classes, the more area enclosed by the AUC. The AUC provides a quantifiable measure in order to differentiate and rank the effectiveness of different classifiers [8].

Despite the strength that ROC and AUC have given to the machine learning community, there are still several cases in which they fail as an evaluation measure. One example is cost, as explained by Drummond and Holte in 2006. The ROC curve does not take into effect the relative "costs" of misclassifications (i.e., the difference between a false positive and a false negative for the application). A cost curve evaluates the normalized cost of misclassification for a particular classifier using a varying range of percent positive example. Positive slopes on the cost curve graph are equivalent to when false positive rate is less than the false negative rate, and negative slopes, when false positive rate is greater. Depending on the application and whether or not false positives or false negatives "cost" more, one can assess the effect of misclassifications for a given classifier. Likewise, the convex hull beneath the cost curve can be made minimal to minimize cost [2]. In recent research, the Brier Curve has been of particular interest as a cost curve that works explicitly with probabilistic labels and has an area underneath, the Brier Score, analogous to AUC [9]. Despite the possibility of incorporating cost curves with ROC to further develop understanding of a classifiers performance, another key issue that arises is that these measures are only limited to a binary class system [2]. Furthermore, a previous expansion of the ROC to multiclass systems was done by Hand and Till in 2001 [10]. Nevertheless this was done in pairwise comparison, with only one class compared to all others.

3 Methodology

3.1 Probabilistic Distributions

Objects classified by a group of people may end up with some disagreement in the class or ranking assigned. This is most pronounced when the experts don't have access to what the others decided, as for example, getting a second opinion from another doctor, or a panel of Olympic judges each providing a different score for a gymnast, based on their own internal scoring metric. Because each opinion is valid and useful, even though different, by assigning probabilities based on the number of experts in agreement, we can produce a probability distribution showing the most likely true case. The goal of a classifier is then to attempt to match this varied response, so as to simulate a panel of experts, and to show multiple possibilities in addition to the highest one. An example of two distinct, multiclass probabilistic distributions can be seen in Figure 2 below. The distance threshold curve that we are proposing in this paper will allow us to evaluate and quantify the differences between such distributions.

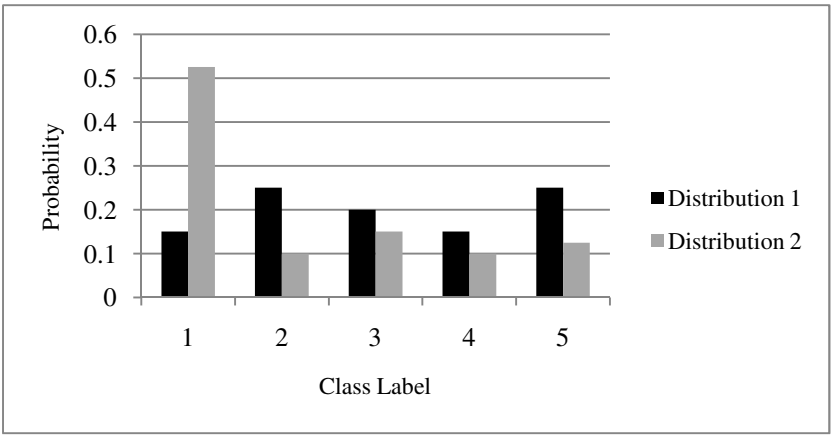

Fig. 2. Example of two distinct probabilistic distributions

3.2 Data

The Lung Image Database Consortium (LIDC) dataset consists of lung nodules rated by up to four radiologists for a variety of semantic characteristics; 64 image features were extracted plus the z-position from the nodule slice images [3]. We then used the data for each image slice, and grouped them by nodule. We filtered by selecting nodules that had been labeled by four radiologists, and then selected two slices, that were located at roughly 33% and 67% through the nodule based on the Z-position value. The slice selection was done programmatically, so as to be able to select multiple images evenly distributed throughout the nodule. We tried different values for the number of slices, and found that two provided a good mix of accuracy and speed. By concatenating the features from the second selected slice to the first, we doubled the number of features from 65 to 130. Two additional features were added based on the entire set of slices for the nodule, Height, and Volume, where height is the distance from the smallest Z-value to the largest, and Volume is the area of each slice multiplied by the thickness, and then summed across all slices. Using this, we were able to produce a data set of 830 out of 2660 total nodules, each with 132 features and characteristic labels from all four radiologists. For our paper, we focused on the semantic characteristic of malignancy, which is scaled on an ordinal range of 1-5, with 1 being the "most-likely benign" and 5 being the "most-likely malignant".

3.3 Classifiers

This data set was then run through a K-Nearest Neighbor classifier, using different values of K. We used a 90%-10% split, where every 10th nodule was set aside for testing purposes, and the other nine were used for training. We varied our number of nearest neighbors from K=2 to K=9 and found that using 5 nearest neighbors was optimum by using the City Block Distance to compare the output results to the radiologist labels and find the difference. From there, we proceeded to use more

sophisticated analysis methods to study the results. The classifier took in multiple labels for each nodule, and produced multiple probabilities as the output. An example is found in Table 2:

Table 2. Sample distribution of radiologist and classifier-predicted probabilistic distributions. In this case, the nodule had been labeled "2" by one radiologist, "3" by two radiologists, and "4" by one radiologist.

Label	1	2	3	4	5
Radiologists Values (A_i)	0%	25%	50%	25%	0%
Predicted Values (B_i)	0%	21.4%	50.0%	25.0%	3.6%

The K-nearest neighbor algorithm sums the Radiologists' Values, A_i, of the K-nearest neighbors (found using the K smallest values of the L_1 norm), and divides by K, to produce the Predicted Values, B_i. The results of the training set are artificially higher than expected due to the fact that for a nodule in the training set, there is a guarantee that one of the nearest neighbors will be the exact nodule being examined, so the results from the testing set are much more useful for determining how well the classifier performed.

In a binary K-NN classifier, the last step is to select the class that is the most frequent, and declare the test case to be of that type. In our case, we take the average probabilities for each type, thus resulting in a distribution similar to B_i in Table 2. A case for probabilistic inputs for multiclass K-NN Classifiers has previously been made in 2009 by Jain and Kapoor [11].

3.4 Evaluation Measures

Accuracy. In order to determine the accuracy of each slice prediction, each probabilistic distribution of values needed to be converted into a single, discrete label. To do so, each distribution by radiologists and predictions were viewed separately and the largest probability was found. The class label index at this case (i.e. 1,2,3,4 or 5), was then assigned as the single label for the slice. In the case that there were two or more probabilities that were equally high, the higher index value was chosen as the label. This was done to err on the side of caution for our malignancy classifier, where a label of "5" is considered the "most-likely malignant".

ROC and AUC. To perform the ROC analysis of the data, a separate conversion of the probabilistic distributions was required from a multiclass to binary array. In order to do so, the five labels had to be isolated into positive and negative classes. Labels 1 and 2 were assigned to the negative class and labels 4 and 5 to the positive. Label 3 is defined as "unknown" malignancy by the LIDC, so was subjected to three conditions:

3 in the positive class (AUC 3+), 3 in the negative class (AUC 3-), or 3 removed from classification evaluation (AUC 3out). It is important to acknowledge that it is conservative to place 3 in the positive class label due to clinical relevancy of the data: in terms of malignancy, it is more appropriate to overestimate than underestimate.

Next positive and negative class probabilities were summed into a binary array of distributions for both radiologists' and predicted label. In both cases, the larger of the two probabilities received the label for that particular slice. The positive class was selected in the case that the two probabilities were 0.5. From there, a simple ROC curve was constructed using the radiologists' labels and the positive probabilities of the predicted labels with a threshold range of 0 to 1 in 0.05 increments. The area under the curve was then extracted using a trapezoidal approximation of the integral.

AUC_{dt}. The Area Under the Distance Threshold Curve was constructed using a variety of distance metrics, including the City Block, Jeffrey Divergence, and Earth Mover's Distance. For any distance measure (d), the AUC_{dt} for instances (S) is given by [3]:

$$AUC_{dt} = \int_0^1 \frac{\sum_{j=1}^{|S|} d(j) \leq x}{|S|} \, dx \tag{4}$$

City Block Distance. The City Block Distance, a member of the Minkowski Distances for p=1 [12]:

$$d_{CB}(A, B) = \sum_{i=1}^{n} (|A_i - B_i|^P)^{\frac{1}{P}} \tag{5}$$

was calculated between each radiologist's probability (A_i) and predicted probability (B_i) for a given label (i=1:5), summed across all labels, and then stored into a vector (d_{CB}) of distances for all nodules.

Jeffrey Divergence. The Jeffrey Divergence was calculated and stored into a similar vector of distances (d_{JD}) with the following algorithm [12]:

$$d_{JD}(A, B) = \sum_{i=1}^{n} [A_i \log\left(\frac{A_i}{\frac{A_i + B_i}{2}}\right) + B_i \log\left(\frac{B_i}{\frac{A_i + B_i}{2}}\right)] \tag{6}$$

Earth Mover's Distance. Earth mover's distance is a measure of the amount of change required to move between two different sets of data, assuming that there is a value to each location as well as the value stored in that location [3]. With the previous distance metrics, there is no difference between a nodule that should be a 5 being incorrectly rated as a 1, and being incorrectly as a 4. In many cases, where the classes are part of a scale, such as a five star rating for restaurants, or the severity of a tumor, there is a large difference between those two types of errors. This method takes that into account, so that a rating of 1 is seen as much worse than a rating of 4, in the case that the real value is 5.

$$EMD_0 = 0 \tag{7}$$

$$EMD_{i+1} = (B_i + EMD_i) - A_i \tag{8}$$

$$d_{EMD}(A,B) = \sum_{i=0}^{N+1} |EMD_i| \tag{9}$$

Weighted Distance. In terms of the clinical application of our classifier, the relative cost between labels was explored through the idea of "weighting" each label. It has already been mentioned that as the label for a nodule's malignancy increases, so does the clinical severity of that label. Therefore, weighted distances for the City Block and Jeffrey Divergence distance metrics were introduced as well. The weighting used was to multiply the difference for each rating by the value of the rating. In other words, for formulas 5 and 6, each probability (A_i and B_i) was multiplied by the label (i). This algorithm puts additional cost weight on the higher malignancy labels, with each label increasing in severity by a factor of one. Nevertheless, this is only a modest weighting assumption as such levels of clinical severity are difficult to quantify.

Construction of the Distance Threshold Curve. After all distance distributions vectors were created, their values were normalized between 0 to 1 for that particular distance metric. For each distance measure, a cumulative histogram was created of the varying distance values for threshold values of 0 to 1 with 0.05 increments and reported the percent frequency of instances for each threshold value (Figure 3). After the curve was plotted, the area beneath the curve was calculated using the trapezoidal integral approximation.

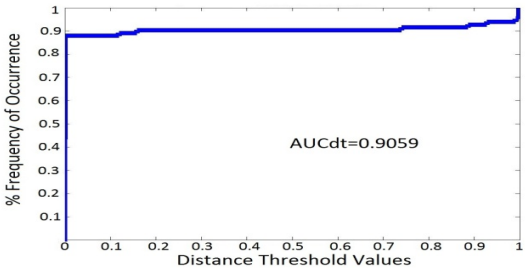

Fig. 3. Example of distance threshold curve constructed with data having 90% agreement. The AUC_{dt} is 0.9059.

In the ideal case of perfect classification the two distributions would be identical and have no difference, producing an AUC_{dt} with 100% frequency of cases for all distance thresholds and an area of 1, similar to the perfect classification instance of ROC curve for the binary case. Similarly, the worst classification gives an area of 0.

3.5 Random Classification of Evaluation Measures

One of the strengths in using accuracy and AUC for ROC evaluation measures is that they have known theoretical behaviors for random classification—that is to say, when a classifier arbitrarily assigns information. In our 5 label multiclass case, accuracy should be roughly 20% for random performance. Likewise, the binary

positive/negative case induced with traditional ROC produces a diagonal line of slope 1 and has an AUC of 0.5 with random classification. However, in the new AUC_{dt} for the multiclass probabilistic case, the random performance was unknown and unique for the various distance metrics.

In order to quantify AUC_{dt} for the arbitrary case, random data was created to test this new measure as well as AUC and accuracy. First, we generated a set of 5000 random probability distributions for Labels 1-5 with a total sum across each distribution of 1. Then we simulated random radiologists for 5000 distributions. The probabilities available in the radiologists' distributions were limited to 0, 0.25, 0.5, 0.75, and 1 (as they would when given deterministic labels). The values of accuracy, AUC, and AUC_{dt} mean and standard deviation of the 50 random radiologist sets with the random predictions are show in Table 3.

Table 3. Mean and standard deviation of evaluation measures with sets of 4 random radiologists and random predictions for 5000 values

Evaluation Measure	Mean	Standard Deviation
Accuracy	0.1999	0.0059
AUC (3+)	0.5080	0.0063
AUC (3-)	0.5055	0.0055
AUC (3out)	0.5054	0.0049
AUC_{dt} CB	0.5391	0.0023
AUC_{dt} JD	0.6891	0.0022
AUC_{dt} EMD	0.7948	0.0015
AUC_{dt} Wtd. CB	0.8374	0.0019
AUC_{dt} Wtd. JD	0.6891	0.0022

It is important to notice that the values for AUC_{dt} with random performance are not only different for a given distance metric (CB, JD, EMD, Wtd. CB, or Wtd. JD), they are all higher than 0.5 in the binary case with AUC for ROC. These values are important to have in evaluating with AUC_{dt} in order to ensure that the classifier is performing better than random in the probabilistic multiclass case.

Next, in order to better understand how to quantify AUC_{dt}, the random classification data was iteratively changed to have varying amounts of agreement between the predictions and the actual radiologists' labels. As there were 5000 random predictions labels and only 747 actual nodules in the training set that were labeled, 4 samples of the 5000 random predictions were taken for each agreement level and the average values were reported. The agreement was increased by 10% from 0 to 90% and then also calculated at 95% and 99%. The evaluation metric values are shown in Table 4.

The information provided regarding random performance in Table 4 allows a certain value of AUC_{dt} to be assessed in terms of similarity between the classified predictions and the actual label values. It is a baseline for which our classifier's performance can be assessed. For example, if our classifier returns an AUC_{dt} value of 0.93 using Jeffrey Divergence, we know that our classifier would be classifying with between 80% and 90% agreement to the radiologists' labels.

Table 4. Performance of AUC_{dt} for all distance metrics comparing actual radiologists' label distributions and random predictions. The random prediction distributions were incrementally changed to have various percentages of agreement with the actual radiologists' labels.

% Agreement	AUC_{dt} CB	AUC_{dt} JD	AUC_{dt} EMD	AUC_{dt} Wtd. CB	AUC_{dt} Wtd. JD
0	0.5225	0.6725	0.7789	0.8220	0.6759
10	0.4769	0.6251	0.7530	0.8130	0.6290
20	0.5484	0.6750	0.7856	0.8410	0.6794
30	0.5805	0.6974	0.7967	0.8443	0.7012
40	0.6336	0.7379	0.8256	0.8683	0.7422
50	0.7350	0.8111	0.8745	0.9061	0.8139
60	0.7563	0.8256	0.8843	0.9087	0.8270
70	0.8269	0.8737	0.9172	0.9383	0.8783
80	0.8684	0.9031	0.9357	0.9497	0.9035
90	0.9433	0.9592	0.9713	0.9771	0.9601
95	0.9699	0.9793	0.9866	0.9909	0.9788
99	0.9946	0.9957	0.9976	0.9984	0.9962

4 Results

The classifiers developed for the semantic characteristic of malignancy were initially evaluated with binary evaluation metrics: accuracy and AUC (3 positive). The same classifier was also evaluated using the newly-introduced metrics for the area under the distance threshold curve: City Block Difference (simple and weighted), Jeffrey Divergence (simple and weighted), and Earth Mover's Distance. For all evaluation measures, a value as close to 1 is desirable. The results of the binary and multiclass measures for the best K-Nearest Neighbor classifier (K=5) are show in Table 5. Figures 4-5 provide the visual representation of all binary and multiclass evaluation metrics.

Table 5. Accuracy, AUC, and Area under the distance threshold curve (AUCdt) for three distance metrics: City Block Distance, Jeffrey Divergence, and Earth Mover's Distance, unweighted and weighted; 5NN Classifier of Malignancy.

Classifier	Accuracy	AUC (3+)	AUC_{dt} CB	AUC_{dt} JD	AUC_{dt} EMD	AUC_{dt} Wtd. CB	AUC_{dt} Wtd. JD
5-NN Training	0.6760	0.9438	0.7321	0.8620	0.8974	0.9187	0.8625
5-NN Testing	0.5542	0.9286	0.6753	0.8141	0.8694	0.8935	0.8144

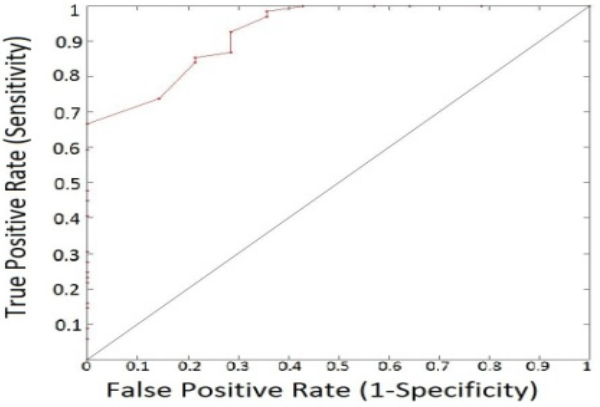

Fig. 4. ROC Curve for 5NN Test Set

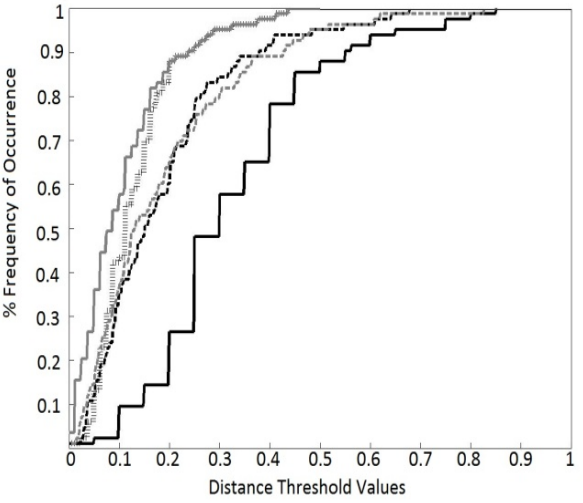

Fig. 5. Distance threshold curve for 5NN Test set—black solid: CBD, black thick dash: JD, black thin dash: EMD, grey solid: weighted CBD, and grey thick dash: weighted JD

5 Discussion

From Table 5 find that the accuracy achieved by the 5NN classifier was 67.60% for training and 55.42% for testing sets. In comparing to the random performance of accuracy in Table 3 (~20%) This low accuracy can namely be attributed to the conversion of our 5-class probabilistic distributions to a single, deterministic label. These results attest that accuracy is not an appropriate measure for evaluating a multiclass case.

The results for the area under the ROC curve in Table 5 show clearly that the 5-Nearest Neighbor classifier is performing well for the binary case: for all definitions of label "3", the AUC has an area above 0.9. However, these values of AUC are different which confirms that AUC is dependent on how the two classes are defined. While our classifier can correctly distinguishing between two distinct classes, we still don't receive information about how well the classifier discriminates single labels.

In regards to our various distance measures for the areas under the distance threshold curve found in Table 5, we find that all of our measures exceed their random performance values from Table 3. Furthermore we note that the weighted JD value of the AUCdt only increased by 0.0003 from its normal JD value. Jeffrey Divergence likely had very little change because the algorithm that calculates the distance is created using a logarithm, where single digit degrees contribute to fewer changes. This suggests that weighting JD is probably not necessary for future applications.

We also can use the values obtained in Table 4 to quantify at about which percentage level of agreement our classifier predicted compared to radiologists' labels. This is done for the different distances used in AUCdt testing sets as follows: for CB, 40-50%, for JD, 50-60%, for EMD, 40-50%, for Wtd. Cb, 40-50%, and for Wtd. JD 50-60%. It is important to note that the range of agreement for all distance measures is around 40-60%. This allows us to argue that our 5-NN classifier is able to match the radiologists' distributions at about 40-60% similarity. Furthermore, the consistency within the AUCdt validates it as a viable measure for the probabilistic, multiclass case.

6 Conclusion

Throughout this paper we have discussed the use and meaning behind traditional, binary measures for evaluation of classifier performance. Our results using a K-Nearest Neighbor classifier with K=5 for malignancy from lung nodule image data from the LIDC database confirm that AUC is a more-effective evaluation tool compared to accuracy. We have also emphasized the need for an evaluation metric for probabilistic, multiclass cases, examining the area under the distance threshold curve. We constructed AUC_{dt} using three distance metrics: City Block Distance, Jeffrey Divergence, and Earth Mover's Distance along with Weighted JD and CB. We have found the random performance values for all distance metrics and showed that Jeffrey Divergence will produce a similar AUC_{dt} when it is weighted and when it is not. This also gives us a comparison to AUC and AUC_{dt} in that they both have known random classification and a perfect classification value of 1.00. We also found that regardless of the distance measure, the AUC_{dt} consistently evaluated our classifier performance as matching the radiologists' labels to about 40-50%. This confirms that AUC_{dt} can be used as an evaluation measure. Furthermore, AUC_{dt} presents an advantage for the multiclass case: ability to distinguish between misclassification of higher or lower labels by using weighted differences.

The use of the AUC_{dt} is especially valuable for the case of medical image annotation. Distinguishing between five distinct semantic classes is essential in radiology,

where the difference between a malignancy ranking of "3" and "4" could be whether to do a lung tissue biopsy or not [14]. Because AUC_{dt} maintains the integrity of each distinct class, it is directly applicable to multiclass classification problems such as those presented within the LIDC.

In the future we will investigate how different distribution metrics can be selected based on the application purposes of the classifier itself; that there is no "perfect" evaluation measure for all cases. We will also explore how viewing this problem as an ordinal regression problem (instead of discrete, deterministic labels) will affect the area under the distance threshold curve. In this case, Earth Mover's Distance will likely be an appropriate measure as it looks at distance amongst classes which is also important in ordinal regression.

Acknowledgements. This work was supported by the National Science Foundation under Grant No. 1062909. We would also like to thank Dimitry Zinovev of DePaul University's College of Computing and Digital Media for providing us with the LIDC image features for each nodule.

References

1. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30, 1145–1159 (1997), doi:10.1016/j.bbr.2011.03.031
2. Drummond, C., Holte, R.: Cost curves: an improved method for visualizing classifier performance. Mach. Learn. 65, 95–130 (2006)
3. Zinovev, D., Furst, J., Raicu, D.: Building an ensemble of probabilistic classifiers for lung nodule interpretation. In: Tenth Intern. Conferen. on Mach. Learn. and App (ICMLA 2011), pp. 151–167. IEEE Press (December 2011), doi:10.1109/ICMLA.2011.44
4. Amor, N.B., Benferhat, S., Elouedi, Z.: Information-based evaluation functions for probabilistic classifiers. In: Eleventh Internat. Conferen. on Infor. Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2006), pp. 428–433 (July 2006)
5. Ling, C.X., Huang, J., Zhang, H.: AUC: a statistically consistent and more discriminating measure than accuracy. In: Proc. of Eighteenth Internat. Conf. on Artifical Intelligence (IJCAI 2003), pp. 519–526 (August 2003)
6. Provost, F., Fawcett, T.: Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In: Proc. Third Internat. Conf. on Knowledge Discovery and Data Mining (KDD 1997), pp. 43–48. AAAI Press (August 1997)
7. Fawcett, T.: ROC graphs: notes and practical considerations for data mining. Technical report, HPL-2003-4
8. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. 27, 861–874 (2006)
9. Hérnandez-Orallo, J., Flach, P., Ferri, C.: Brier curves: a new cost-based visualization of classifier performance. In: Proc. Twenty-Eighth Internat. Conf. on Mach. Learn (ICML 2011), pp. 585–592 (June 2011)
10. Hand, D.J., Till, R.T.: A simple generalization of the area under the ROC curve for multiple class classification problems. Machine Learning 45, 171–186 (2001)
11. Jain, P., Kapoor, A.: Active learning for large multi-class problems. Comp. Vision and Pattern Recogn (CVPR 2009), 762–769 (June 2009)

12. Liu, H., et al.: Comparing dissimilarity measures for content-based image retrieval. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 44–50. Springer, Heidelberg (2008)
13. Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. In: Sixth Internat. Conf. Comp. Vis. (ICCV 1998), pp. 59–66. IEEE Press (January 1998), doi:10.1109/ICCV.19
14. Raicu, D.S., Varutbangkul, E., Furst, J.D., Armato III, S.G.: Modeling semantics from image data: opportunities from LIDC. Internat. Jour. of Biomed. Eng. and Tech. 3(30:1-2), 83–113 (2009)