IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Special Issue-1 E-ISSN: 2347-2693

A Study on Benchmarking Parameters for Intelligent Systems

Rajesh Misir[1*]

Department of Computer Science,
Vidyasagar University,
Midnapur - 721102.
[1]rajeshmisir@gmail.com

Malay Mitra[2] **and R. K. Samanta**[3]
Dept. of Computer Science & Application,
University of North Bengal, Darjeeling- 734013.
[2]malay.mitra68@gmail.com, [3]rksamantark@gmail.com

***Abstract*- Intelligent automated decision support systems are now found to be very much useful in various fields. In bioinformatics and machine learning in general, there is a large variation in the predictive measures that are used to evaluate intelligent systems. If we do not assess the accuracy of model's prediction, a vital step in model development, its application will have little merit. This work critically discusses different approaches to assess predictive performance and various test statistics. Choice of assessing strategy or validation for a specific application helps in determining the suitability of the model and in comparing the performances of different modeling techniques. The purpose of this paper is to serve as an introduction to various important benchmarking parameters and as a guide for using them in research.**

Keywords: Predictive performance, Confusion Matrix, Receiver operating characteristic (ROC), Akaike information criteria (AIC), Kappa statistic, Lift, Cumulative gain, Probability Threshold

I. INTRODUCTION

Intelligence can be defined in different ways such as in terms of one's capacity of learning, understanding, planning, communication, problem solving, reasoning etc. A system showing intelligence is an intelligent system. Most of the techniques for building intelligent models stem from the discipline termed artificial intelligence (AI). Since the early 70's, one of the prime research fields in computer science is to make a computer system intelligent under the broad discipline AI.

For judging the quality and quantity of human intelligence, IQ (Intelligent Quotient) is used as one of the benchmarking parameters. Likewise, EQ (Efficiency Quotient) for intelligent systems can be measured in a number of ways using different test statistics and data mining parameters. This study will reveal some such parameters with their relevant application areas and also their limitations keeping in mind the axiom of data mining that each data set is unique. This intends to help apply suitable benchmarking parameter(s) for a problem domain.

Section 2 to section 8 explains different performance prediction measures; their strengths and limitations. Lastly, conclusions are summarized in section 9.

II. CONFUSION MATRIX

Accuracy of a classification model can be assessed by a confusion matrix. It summarizes predictive performance. In general, a confusion matrix is an n-dimensional square matrix, where n is the number of distinct target values. For a binary classification model, it is a two dimensional square matrix. It records the frequencies of each of the four possible types of prediction from analysis of test data:

i) true positive i.e. positive cases in the test data with predicted probabilities greater than or equal to the probability threshold (correctly predicted): **TP**

ii) false positive i.e. negative cases in the test data with predicted probabilities greater than or equal to the probability threshold (incorrectly predicted): **FP**
iii) true negative i.e. negative cases in the test data with predicted probabilities less than to the probability threshold (correctly predicted): **TN**
iv) false negative i.e. positive cases in the test data with predicted probabilities less than to the probability threshold (incorrectly predicted): **FN**

Figure 1 shows a confusion matrix which depicts prediction of instances for a binary classification model.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 1: Confusion Matrix

Most commonly used test statistics derived from confusion matrix are given in Table 1 [1].

Table 1: Most commonly used test statistics derived from confusion matrix

Measure	**Formula**	**Meaning**
Precision	TP/(TP+FP)	Percentage of positive predictions those are correct.
Recall/Sensitivity	TP/(TP+FN)	Percentage of positive labeled instances that were predicted as positive.
Specificity	TN/(TN+FP)	Percentage of negative labeled instances that were predicted as negative.
Accuracy	(TP+TN)/(TP+FP+TN+FN)	Percentage of correct predictions.

The name confusion matrix stems from the fact that it points out where the model gets confused i.e. makes incorrect prediction. It can be used not only to judge the classification performance, but also to judge the misclassification cost incurred by specifying the cost of right and wrong classification. Different problem domains need different measures to summarize prediction quality. For example-

1) In a data set of 5000 instances, where only 50 instances are labeled as positive and the model predicts "Negative" for all instances. Here accuracy is 99% and specificity is 100%, but sensitivity is 0% indicating problem in the model.
2) If in the above case the model predicts "Positive" for every instance, the sensitivity, specificity and accuracy will be 100%, 0% and 1%, respectively. Hence specificity and accuracy reflect that the classifier is problematic.
3) Consider another case, wherein out of a data set of 5000 instances, 4950 instances are recorded as "Positive". If the model predicts "Positive" for every case, the sensitivity (100%) and accuracy (99%) reflect that the model performs well, but the specificity (0%) indicates problem in the model.
4) If in the above case the model predicts "Negative" for every case the accuracy (1%) and sensitivity (0%) reflect that the model is problematic though the specificity is 100%.

So, it can be proposed that without any knowledge of the distribution of data all the three measures are of equal importance to predict the performance of a model. It can also be shown that these three measures only are not sufficient to predict. Since in Biological applications the majority of the examples are negative [2], precision and sensitivity play an important role to evaluate the model.

III. AREA UNDER ROC CURVE

A receiver operating characteristic (ROC) or simply ROC curve, a part of a field called "signal detection theory", is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) against the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings. The Receiver Operating Characteristic (ROC) curve is a technique that is widely used in machine learning experiments representing a graphical plot that summarizes how a classification system performs and allows us to compare the performances of different classifiers. The area under the ROC curve is a measure of overall performance of a classification model. Generally, the higher the area under the ROC curve, the better the model performance. It is way to compare classification model quality by determining false positive and true positive rates at different probability thresholds. Besides model selection, the ROC also helps to determine a threshold value to achieve an acceptable trade-off between hit (true positives) rate and false alarm (false positives) rate. By selecting a point on the curve for a given model a given trade-off is achieved. This threshold can then be used as a post-processing parameter for achieving the desired performance with respect to the error rates. ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems [3]. The medical decision making community has an extensive literature on the use of ROC graphs for diagnostic testing [4]. The true positive rate (TPR), also called sensitivity of a classifier is evaluated as:

$$TPR=TP/(TP+FN)$$

The false positive rate (FPR) of classifier is estimated as:

$$FPR = FP/(FP+TN)=1- \text{specificity}$$

ROC graphs are two dimensional plots in which TPR is plotted on the Y-axis and FPR, i.e. 1-specificity is plotted on the X-axis. A discrete classifier produces a pair (FPR, TPR) that corresponds to a single point in ROC space. Fig. 2 shows ROC graph with five classifiers labeled A through E [5].

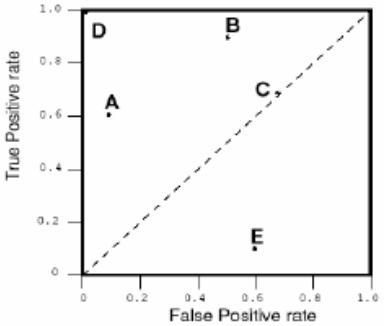

Figure 2: Classifier performance- Points in ROC Space

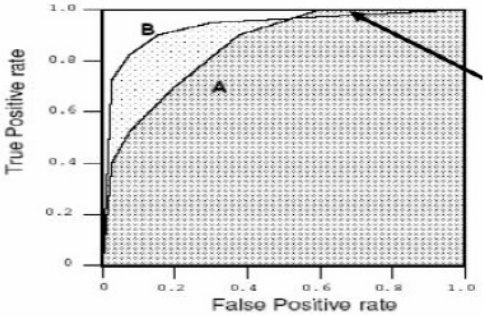

Figure 3: Comparison of performance by AUC of classifier A and B [5]

One can easily interpret different points in the ROC space. For example the point (0,0) indicates no false positive errors, but gains no true positives. The point (0,1) represents perfect classification. Informally, a point in ROC space is better than another if it is to the northwest of the first. The ROC space has two axes, each having a maximum value of 1. ROC curve is defined by plotting TPR against FPR across the range of possible thresholds. Each threshold value produces a different point in ROC space. ROC curve can be obtained by tracing these points. The area under ROC curve measures discrimination, that is, the ability of the test to correctly classify those with and without a specific property.

An ROC curve represents following things-

a) It shows the tradeoff between sensitivity and specificity.

b) The closer the curve follows the left hand border and then the top border of the ROC space, the more accurate the test.

c) The closure the curve comes to the 45˚ diagonal of the ROC space (1:1 line), the less accurate the test.

d) The area under the curve (AUC) is a measure of classification accuracy.

In order to summarize predictive performance across the full range of thresholds we can measure the area under ROC curve (AUC), expressed as a proportion of the total area of the square defined by the axes [3]. The AUC ranges from 0.5 for models that are no better than random to 1.0 for models with perfect predictive ability. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system, which is as follows–

0.90-1.00 represents Excellent (A)
0.80-0.90 represents Good (B)
0.70-0.80 represents Fair (C)
0.60-0.70 represents Poor (D)
0.50-0.60 represents Fail (F)

Limitations of ROC – It is possible for a high AUC classifier to perform worse in a specific region of ROC space than a low AUC classifier. In Figure 3 the high AUC classifier B performs worse than the low AUC classifier A for FPR > 0.6.

IV. KAPPA STATISTIC

The medical diagnostic test results will be of little use, if the people who interpret the test cannot agree on the interpretation. Kappa statistic, suggested by Cohen in 1960 [6], is a generic term for several similar measures of agreement used with categorical data. Typically it is used in assessing the degree to which two or more raters, examining the same data, agree when it comes to assigning the data to categories. Kappa might be used to assess the extent to which radiology analysis of an X-ray, computer analysis of the same X-ray and biopsy agree in labeling a growth as malignant or benign. In recent years, the Kappa coefficient of agreement has become the de facto for evaluating inter-coder agreement for tagging tasks.

For example, let us consider two observers, denoted by rater A and rater B, classify 100 subjects into one of two possible classes, labeled as 1 and 2. The Kappa value is calculated based on the difference between the observed agreement (how much agreement is actually present) and the expected agreement (how much agreement would be expected to be present by chance alone). The data layout is given in Table 2. The observed agreement (P_o) is the percentage of all frequencies for which two raters agree, i.e. (a + b) / (a + b + c + d). In the given example it is (25+60)/100 = 0.85.

Table 2: Data Layout

		Rater A		
Rater B	1	1	2	Total
		25(a)	10(b)	35(m_1)
	2	5(c)	60(d)	65(m_0)
		30(n_1)	70(n_0)	100(n)

(a) and (d) denote the number of times the two raters agree while (b) and (c) denote the number of times the two raters disagree.

Expected agreement (P_e) is evaluated from the formula:

$$P_e = [(n_1/n) * (m_1/n)] + [(n_0/n) * (m_0/n)]$$

For the above example the value of P_e is:

$$P_e = [(30/100) * (35/100)] + [(70/100) * (65/100)]$$
$$= [0.3 * 0.35] + [0.7 * 0.65] = 0.56$$

Kappa, K, is defined as:

$$K=(P_o-P_e)/(1-P_e) = (0.85 – 0.56) / (1 – 0.56) = 0.66$$

The value of Kappa lies between -1 to 1, where perfect agreement would equate to a Kappa of 1, chance agreement would equate to a Kappa of 0 and negative values indicate potential systematic disagreement between the observers. A common interpretation of Kappa is as given in Table 3 [7].

Table 3: Interpretation of Kappa

Kappa	Agreement
<0	Less than chance agreement
0.01 to 0.20	Slight agreement
0.21 to 0.40	Fair agreement
0.41 to 0.60	Moderate agreement
0.61 to 0.80	Substantial agreement
0.81 to 0.99	Almost perfect agreement

Limitations of Kappa: Let us consider 100 subjects of two experiments are classified by Rater A and Rater B into one of two possible categories, labeled as 1 and 2. Table 4 and Table 5 show the outcomes of the experiments.

Table 4: Outcomes of Experiment 1

Rater B	Rater A		
	1	2	Total
1	25	8	33
2	7	60	67
Total	32	68	100

Table 5: Outcomes of Experiment 2

Rater B	Rater A		
	1	2	Total
1	81	9	90
2	6	4	10
Total	87	13	100

In both cases the observed agreement is 0.85. So, one may expect high inter-rater reliability for both cases (i.e. high value of Kappa). Cohen's Kappa for these two cases are 0.66 for Experiment 1 and 0.265 for Experiment 2. Thus Kappa statistic for Experiment 2 reflects the low level of agreement between the raters. It is very difficult to explain why raters have substantially high level of agreement for Experiment 1, while a fairly low agreement for Experiment 2. This paradox shows that there are serious conceptual flaws in Kappa statistics.

V. AKAIKE INFORMATION CRITERIA (AIC)

Akaike [8] adopted the Kullback-Leibler definition of information, I(f;g), as a natural measure of discrepancy, or asymmetrical distance, between a true model, f(y), and a proposed model, g(y|β), where β is a vector of parameters. Based on large sample theory, Akaike derived an estimator for I(f;g) of the general form:

$$AIC_m = 2K_m - 2Ln(L_m)$$

where L_m is the sample log-likelihood for the m^{th} of M alternative models and K_m is the number of independent parameters estimated for the m^{th} model. AIC provides a means for selecting a model. For a given set of data, it is a measure of the relative quality of a statistical model. Using AIC, one can compare normal models, gamma models, lognormal models, square root normal models etc. Various features of AIC are as follows:

1) AIC does not provide a test for a model. It is used to select the best among two or more competing models.
2) A min (AIC) strategy [7] is used for selecting a model from a set of candidate models. Given a set of candidate models, the perfect model is the one with the minimum AIC value, specific to given data set.
3) AIC can compare models with different error distribution.
4) A potential danger arises in count regression models, where the natural data generating mechanisms might be Poisson, negative binomial, etc. but we wish to include in this mix probability models that are more suitable for continuous responses: normal, lognormal, etc.

A popular alternative to AIC presented by Akaike [9] and Schwarz [10] is Bayesian Information Criterion (BIC). In comparison with BIC, AIC is asymptotically optimal in selecting the model with the least mean squared error, under the assumption that the exact true model is not in the candidate set; BIC is not asymptotically optimal under the assumption. The AIC penalizes the number of parameters less strongly than the BIC. The difference between two BIC estimates may be good approximation to the natural log of the Bayes factor [11].

VI. LIFT

The lift curve helps to select a relatively small number of cases and to get a relatively large portion of the responders. Validation data set, input to construct a lift curve, has been "scored" by appending to each case the estimated probability that it will belong to a given class. The graph is plotted with the cumulative number of cases (in descending order of probability) on the X-axis and the cumulative number of true positives on the Y-axis as shown in fig. 4. A good classifier will give a high lift acting on only a few cases. The lift curve can also be considered as a variation on the receiver operating characteristic (ROC) curve. A lift chart graphically represents the improvement that a mining model provides when compared against a random guess and measures the change in terms of a lift score. By comparing the lift scores for various portions of data set and for different models, one can determine which model is best and which percentage of the cases in the data set would benefit from applying the model's predictions. So it is evident that:

- Lift is a measure of effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.
- Model's performance can be assessed visually from lift charts.
- The better will be the performance of the model if the area between the lift curve and the baseline is greater [12].

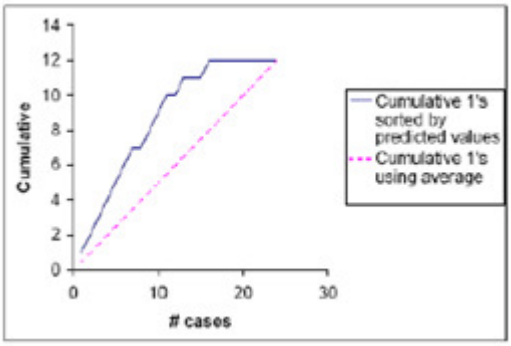

Figure 4: Lift Curve

VII. CUMULATIVE GAIN

It is calculated as follows:

Gain = (Expected response using predictive model) / (Expected response from random mailing)

It is the percentage of positive responses determined by the model across quantiles (In statistics points on a probability distribution function separated by the same fraction of the probability; there is an integrated probability of 1/n between two adjacent n-quantiles) of the applied data. Cases are typically divided into 10 or 100 quantiles against which cumulative gain (and Lift also) is reported. Cumulative gain for a given quantile is the ratio of the cumulative number of positive targets to the total number of positive targets.

VIII. PROBABILITY THRESHOLD

Conversion of continuous model output into binary predictions is useful to predict "Present" or "Absent" by setting a threshold probability value above which the species is predicted to be "Present". Different threshold probability values result in different false positive rates and true positive rates. This approach is not suitable in many circumstances notably when some records are not available [13]. Different methods used to select probability threshold are listed in Table 6.

Table 6. Some published methods for setting threshold of occurrence [14, 15]

Method	Procedure	Species data type	Reference(s)
Fixed value	An arbitrary fixed value (e.g. probability = 0.5)	Presence only	[16, 17]
Lowest predicted value	The lowest predicted value corresponding with an observed occurrence record	Presence only	[18, 19]
Fixed sensitivity	The threshold at which an arbitrary fixed sensitivity is reached (e.g. 0.95, meaning that 95% of observed localities will be included in the prediction)	Presence only	[20]
Sensitivity-specificity equality	The threshold at which sensitivity and specificity are equal	Presence and absence	[20]
Sensitivity-specificity sum maximization	The sum of sensitivity and specificity is maximized	Presence and absence	[21]
Maximize Kappa	The threshold at which Cohen's Kappa statistic is maximized	Presence and absence	[22, 23]
Average probability/suitability	The mean value across model output	Presence only	[24]
Equal prevalence	Species' prevalence (the proportion of presences relative to the number of sites) is maintained the same in the prediction as in the calibration data.	Presence and absence	[24]

If the purpose of modeling is to identify areas within which disturbance may impact a species negatively then the threshold may be set low to identify a larger area of potentially suitable habitat. In contrast, if the model was intended to identify potential introduction or reintroduction sites for an endangered species or species of recreational value, then it would be appropriate to choose a relatively high threshold. Choosing a high threshold reduces the risk of choosing unsuitable sites by identifying those areas with highest suitability [25].

IX. CONCLUSION

From the present study, it is revealed that no single approach from the above can be recommended for all applications. The choice of selecting performance parameters depend on the model and the type of data available. In general, four measures, viz. precision, sensitivity, specificity, classification accuracy i.e. the entire confusion matrix should be recommended as performance prediction parameters without any knowledge of the distribution of data. Area under receiver operating characteristic also plays an important role to depict performance of a model. Other parameters are used to compare models and to enhance the performance of models.

REFERENCES

[1] Z. Lu et al., Predicting Subcellular Localization of Proteins using Machine Learned Classifiers, Bioinformatics, vol 20, issue 4, pp. 547-556, 2004.

[2] R. Eisner et al., Improving Protein Function Prediction using the Hierarchical Structure of the Gene Ontology, IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Nov-2005. --

[3] J. A. Swets, Measuring the accuracy of diagnostic systems. Science, 240, pp. 1285–1293,1988.

[4] Xiao-Hua Zhou and Jaroslaw Harezlak, Comparison of bandwidth selection methods for kernel smoothing of ROC curves, Statist. Med.; 21:2045–2055, 2002.

[5] Tom Fawcett, An introduction to ROC analysis , Pattern Recognition Letters, vol 27, pp. 861-874, 2006.

[6] Jacob Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20 (1), pp. 37–46, 1960.

[7] J.R. Landis., G.G. Koch, The measurement of observer agreement for categorical data, Biometrics, vol 33, no. 1, pp 159-174, 1977.

[8] H. Akaike, Information Theory and An Extension of The Maximum Likelihood Principle. In B.N. Petrov and F. Caske (eds.), Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267-281, 1973.

[9] H. Akaike, A Bayesian Analysis of The Minimum AIC procedure, Ann. Inst. Statist. Math., vol 30, no. 1, pp. 9-14, 1978.

[10] G. Schwarz, Estimating the Dimension of a Model, The Annals of Statistics, Vol. 6, no. 2, pp. 461-464, 1978.

[11] R. E. Kass and L. Wasserman, A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion, J. American Stat. Association, Vol. 90, no. 431, pp. 928-934, 1995.

[12] Tariq Jaffery, Shirley X. Liu, Measuring Campaign Performance by Using Cumulative Gain Lift Chart, SAS Global Forum 2009, Paper 196, 2009.

[13] S. J. Philips et al., Maximum entropy modeling of species geographic distributions, Ecological Modelling, Vol. 190, pp. 231–259, 2006.

[14] C. Liu et al, Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, pp. 385-393, 2005.

[15] R. G. Pearson, Species' Distribution Modeling for Conservation Educators and Practitioners, Synthesis. American Museum of Natural History, 2007. Available at http://ncep.amnh.org.

[16] S. Manel et al., Comparing discriminant analysis ,neural networks and logistic regression for prediction species distributions: a case study with a Himalayan river bird. Ecological Modelling, 120, 337-347, 1999.

[17] M. P. Robertson et al., A PCA-based modeling technique for predicting environmental suitability for organisms from presence records, Diversity and Distributions 7, 15-27, 2001.

[18] R. G. Pearson et al., . Model based uncertainty in species' range prediction, Journal of Biogeography, 33, pp. 1704-1711, 2006.

[19] S. J. Phillips et al., Maximum entropy modeling of species geographic distributions, Ecological Modelling , Vol. 190, pp. 231-259, 2006.

[20] R. G. Pearson et al., . Modelling species distributions in britain: A hierarchical integration of climate and land-cover data, Ecography, 27, 285-298, 2004.

[21] S. Manel et al., Evaluating presences absence models in ecology: the need to account for prevalence, Journal of Applied Ecology, Vol. 38, pp.,921-931, 2001.

[22] B. Huntley et al., Modelling present and potential future ranges of some European higher plants using climate response surfaces. Journal of Biogeography, Vol. 22, pp.967-1001, 1995.

[23] J. Elith et al., Novel methods improve prediction of species' distributions from occurrence data, Ecography, Vol. 29, pp. 129-151, 2006.

[24] J. S. Cramer,. Logit models: from economics and other fields. Cambridge University Press, 2003.

J. Pearce et al., Evaluating the predictive performance of habitat models developed using logistic regression, Ecological Modelling, Vol. 133, pp. 225–245, 2000.