CHAPTER 5

Performance Predic on Parameters

5.1. Introduc on

One's capacity of handling problems, learning, understanding, planning, communication, problem solving, and reasoning are measure of intelligence. A system that can take decision based on intelligence is an intelligent system. Artificial Intelligence, a branch of computer science deals with the techniques to develop intelligent system. 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.

IQ (Intelligent Quotient) is a parameter for measuring intelligence of human being. Similarly, various prediction parameters and test statistics are used to measure efficiency quotient (EQ) for intelligent models. This chapter presents a study on such type of parameters, their usefulness and limitations, considering the axiom of data mining that each data set is unique. The objective is to find out the suitable performance prediction parameters for a particular problem domain

Predictive performance of classifier should be measured to compare classification models. Both supervised and unsupervised classification model has been trained on training data set. The predictive performance of the trained classifier on the test data set is measure of the performance of the intelligent model on unknown data. Some parameters should be measured to assess the performance of the classifier and its generalization ability, e.g. error rate, accuracy etc.

Different performance prediction measures; their strengths and limitations are explained in this section. Lastly, conclusions are summarized.

5.2. Confusion Matrix

Predictive performance of an intelligent classification model is summarized by

[iii] This chapter is based on the publication made by the author entitled "A Study on Benchmarking Parameters for Intelligent Systems", International Journal of Computer Sciences and Engineering (IJCSE), Vol. 3, Issue 1, pp. 10-17, February, 2015, E-ISSN: 2347 - 2693.

confusion matrix, which is useful while assessing the performance of it. Generally, a confusion matrix is an n-dimensional square matrix, where n is the number of distinct target values. In case of a binary classification model, it is represented by a two dimensional square matrix. Confusion matrix visualizes the performance of a classification model based on a given test data set by representing instances in predicted class in each column and instances in actual class in each row. It records the frequencies of each of the four possible types of prediction from analysis of test data:

i) true positive denoted as **TP**, which represents number of positive cases in the test data with predicted probabilities greater than or equal to the probability threshold i.e. correctly predicted.

ii) false positive denoted as **FP**, which represents number of negative cases in the test data with predicted probabilities greater than or equal to the probability threshold i.e. incorrectly predicted.

iii) true negative denoted by **TN**, which represents number of negative cases in the test data with predicted probabilities less than to the probability threshold i.e. correctly predicted.

iv) false negative denoted by **FN**, which represents number of positive cases in the test data with predicted probabilities less than to the probability threshold i.e. incorrectly predicted.

Figure 5.1 represents a confusion matrix for which the 'class' variable takes only two values i.e. binary classification.

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

Figure 5.1. Confusion Matrix

Most commonly used benchmarking parameters to assess predictive performance of an intelligent system, derived from confusion matrix are given in Table 5.1 [1].

Table 5.1. Most Commonly used Test Sta s cs Derived from Confusion Matrix

Measure	Formula	Meaning
Precision	TP/(TP+FP)	Percentage of correct positive predictions.
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 confusion matrix precisely reflects the fact in which condition a model fails by providing incorrect prediction or getting confused. For this is reason it is called confusion matrix. If the costs of right and wrong classifications are specified, it can also be used to measure the misclassification cost. Decision has been taken to choose performance measure based on problem domains. For example-

1) A data set containing 6000 instances, in which the number of instances labeled as positive is 60. “Negative” prediction of the model for any instance yields 99% accuracy and 100% specificity, but sensitivity in this case is 0%, which indicates 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) Similarly in another case, where 5950 instances of a data set of size 6000 instances are recorded as "Positive", “Positive” prediction for every case of a model results in the sensitivity (100%) and accuracy (99%). From these results one may think that the model is performing well, but the specificity (0%) indicates problem in the model.

4) "Negative" prediction for every case for the above example results the accuracy (1%) and sensitivity (0%) reflect that the model is problematic though the specificity is 100%.

From the above discussions it is evident that unbalanced data, where number of instances of different classes differs greatly, may mislead in predicting the performance of an intelligent system. For this reason it is suggested to consider all the four parameters, derived from the confusion matrix, as all of them are of equal importance while judging a predictive model without any knowledge of the distribution of data. Sensitivity and precision play a vital role to predict the performance of model elated to biological applications as most of the instances are negative [2]. Other benchmarking parameters may have to be considered while assessing the predictive performance of a model.

5.3. Area Under ROC Curve

In “signal detection theory” receiver operating characteristic or ROC curve is used, which is also an important measure to illustrate the performance of a binary classifier graphically. It is a plot of TPR (true positive rate) against FPR (false positive rate) at different threshold settings, where TPR rate of correct positive prediction and FPR represents rate of correct negative prediction. Recently, in machine learning ROC curve is used widely to visualize the performance of a model graphically as well as to compare different models. AUC (area under ROC curve) is an important parameter to assess the overall performance of a classifier. Classification models can be compared by evaluating AUC at various 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. The performance of diagnostic systems can be viewed and analyzed by ROC analysis [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 given in Eq. 1.

$$TPR=TP/(TP+FN) \qquad (1)$$

The false positive rate (FPR) of classifier is estimated as given in Eq. 2.

$$FPR = FP/(FP+TN)=1- \text{specificity} \qquad (2)$$

A two dimensional graph in which X-axis represents FPR i.e. 1 – specificity and Y axis represents TPR is known as ROC curve. A discrete classifier produces a pair (FPR, TPR) that corresponds to a single point in ROC space. Fig. 5.2 shows ROC graph with five classifiers labeled A through E [5].

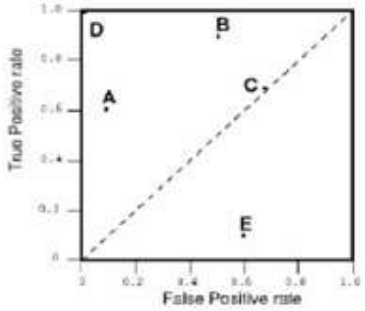

Figure 5.2. Classi er Performance-Points in ROC Space [5]

Interpretation of ROC curve is simple. As for example the point of coordinate (0,0) specifies no false positive errors, but not gaining true positives. The coordinate of perfect classification point is (0,1), which represents no false positive errors and gains 100% sensitivity. Informally, if a point in ROC space is at the northwest of another, the former point represents better performance.

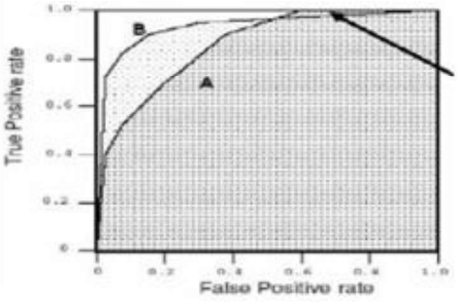

Figure 5.3. Comparison of Performance by AUC of Two Classi ers

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 points out the followings:

a) Balance between sensitivity and specificity is depicted.

b) If the curve follows the path which is initially closer to left hand border and then towards top border, it represents good performance with higher accuracy.

c) A curve closer to 45˚ diagonal line of the ROC space (1:1 line) represents lower performance with less accuracy.

Thus area under ROC curve can be used as a measure to predict performance of an intelligent system. 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 guidance provided by the traditional academic point system to estimate the performance of a model based on AUC are as follows:

0.90-1.00 specifies Excellent (A)

0.80-0.90 specifies Good (B)

0.70-0.80 specifies Fair (C)

0.60-0.70 specifies Poor (D)

0.50-0.60 specifies Fail (F)

5.3.1. Limita ons of ROC Curve

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. Several features of ROC curve are barriers to use it as an only performance predictive measure of classifier. As a discrimination index AUC represents the likelihood that a presence will possess larger predicted value than an absence, regardless how well the prediction fits a set of observations. Thus a well fitted model may lead to poor discrimination power, if the probability of presence is moderately higher than the absence. It is also possible that a poorly

fitted model may cause good discrimination.

Another drawback of ROC curve is that it summarizes predictive performance over the regions of ROC space where one would operate rarely. The extreme left and the right side of the ROC space are valueless as they represent high false negative rates and high false positive rates respectively. The central and left areas of the curve become valueless if we are concentrating on maximizing true positives and avoiding commission error.

As an alternative approach partial ROC has been proposed which was also unable to avoid the drawbacks as mentioned in the literature. Moreover, the spatial distribution of classifier errors are not identified in ROC curves.

5.4. Kappa Sta s c

In health care system more than one person may collect data of clinical laboratory. Agreement among the individuals who collect data should be measured as the medical diagnostic test results will be of little use, if the people who interpret the test cannot agree on the interpretation. To do so well designed research studies involve – training the data collectors to reduce the amount of disagreement and measuring the extent to which the observers agree for the same event. The “inter-rater reliability” is the extent of agreement between the observers.

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-rater agreement for tagging tasks.

Let us consider there are two observers classifying 100 subjects. Assume that the class variable takes only two values, labeled as 1 and 2. Consider Table 5.2 represent the data layout, where a and d denote number of agreements and b and c denote number of disagreement.

Table 5.2. 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)

Let P_o denotes the observed agreement, which is the percentage of all frequencies for which two raters agree, i.e. (a + d) / (a + b + c + d). For the data layout as given in Table 5.2 P_o is (25+60)/100 = 0.85.

Expected agreement (P_e) is evaluated from the Eq. 3.

$$P_e = [(n_1/n) * (m_1/n)] + [(n_0/n) * (m_0/n)] \quad (3)$$

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 5.3. Interpreta on 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

5.4.1. Limita ons of Kappa Sta s c

Let us consider two observers, denoted by Rater A and Rater B, classifying 100 subjects by using two experiments in two classes, labeled as 1 and 2. Experimental results are shown in Table 5.4 and Table 5.5.

Table 5.4. Outcomes of Experiment 1

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

Table 5.5. Outcomes of Experiment 2

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

From the above table it is observed that for both experiments two raters agree in 85 cases i.e. the observed agreement is 0.85. So, one may mislead by expecting high inter-rater reliability for both cases (i.e. high value of Kappa). Cohen's Kappa for Experiment 1 is 0.66, and for Experiment 2 is 0.265. Hence Experiment 2 low level of agreement between the raters. The reason of explaining the given situation is very difficult in this case. There may be serious conceptual flaws in Kappa statistics due to this contradiction.

The assumptions about the rater independence and other factors in Kappa are not well supported, which may lower the estimate of agreement excessively. Thus researchers may have to accept the low Kappa values in their inter-coder reliability studies. In health care system low values of Kappa are not accepted as it may lead to poorer patient outcome [8].

5.5. Akaike Informa on Criteria and Bayesian Informa on Criterion

Akaike [9] 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. The estimator for I(f;g) has been derived by Akaike, based on large sample theory. The derived estimator is of the general form as given in Eq. 4.

$$AIC_m = 2K_m - 2Ln(L_m) \qquad (4)$$

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.

The relative distance (Δ_m) between each model to the truth is a measure in determining the best model (Eq. 5.).

$$\Delta_m = AIC_m - \min AIC \qquad (5)$$

where, min AIC is the minimum value of AIC of the candidate models.

After calculating the relative distances, Akaike weights (w_m) is to be evaluated for each model using the formula as given below:

$$w_m = \frac{\exp(-0.5*\Delta m)}{\sum_{r=1}^{R} \exp(-0.5*\Delta m)} \qquad (6)$$

The denominator gives the sum of exp(-0.5* Δ_m) for candidate models. The general rule of thumb is that the models with Akaike weights within 10% of the highest w_i are in confidence set of candidate models.

Various features of AIC are as follows:

1) AIC is suitable for comparing two or more models to select the best one. It can be used to compare models with different error distribution. It is not used to test performance of a model.

2) For a specific data set a model with minimum AIC is chosen as the best performer from a set of models. This process of selecting model is known as min (AIC) strategy [7].

BIC, presented by Akaike [10] and Schwarz [11], is also used to select model with the lowest BIC value from a set of candidate models and can be think of as an alternative to AIC. Likelihood function is the basis of BIC. Both BIC and AIC solve problems by incorporating a penalty term, which is larger in case of BIC, into the model. BIC is calculated from Eq. 6 as given below.

$$BIC = \ln(n)k - 2\ln(L) \qquad (7)$$

where, L denotes maximized value of likelihood function, n is the number of data points in observed data, and is the number of parameters estimated by the model.

5.5.1. Limita ons of AIC and BIC

When a model, say A, is nested in another model, say B, a small difference in AIC and BIC should not be used as only metric to judge the models. The parameters of interest of models may take the values nearer to boundary of the parameter values

space. This boundary problem may affect the goodness of fit measures, such as AIC and BIC. BIC fails in handling complex collection of models.

5.6. Li

The ratio of percentage of true positive cases made by a predictive model to the percentage of actual positive cases in the test data set is known as lift. If 70% of test data is actually positive and the model predicts 60% as true positive, the value of lift will be 0.8571 (0.60/0.70).

Data set is partitioned into quantiles, which are different portions of data set with equal number of instances. Lift is evaluated against each quantile. The data set is quantiled after scoring. A rank is assigned to each quantile based on probabilities of the positive class prediction. The quantiles are sorted in descending order of rank.

Large portion of responders can be obtained with the help of lift curve. It also helps in selecting small number of cases. Lift curve is constructed based on validation data set, which has been scored by appending estimated probability to each case that it will belong to a given class. The lift curve is obtained by plotting the cumulative number of cases (probability is in descending order) on the X-axis and the cumulative number of true positives on the Y-axis as presented in Figure 5.4. High lift acting on few cases represents a good classifier. Like ROC curve it also depicts the performance of models. Determination of the best model can be made by comparing the lift scores considering various portion of data set for different models. The usefulness of lift curve is as follows:

- It can be used to measure the effectiveness of a predictive system. The evaluation is made by calculating the ratio between the outcomes with and without the predictive system
- It is a pictorial representation of performance of predictive models.
- Greater the area bounded by lift curve and the baseline, the better will be the performance of the predictive model [12].

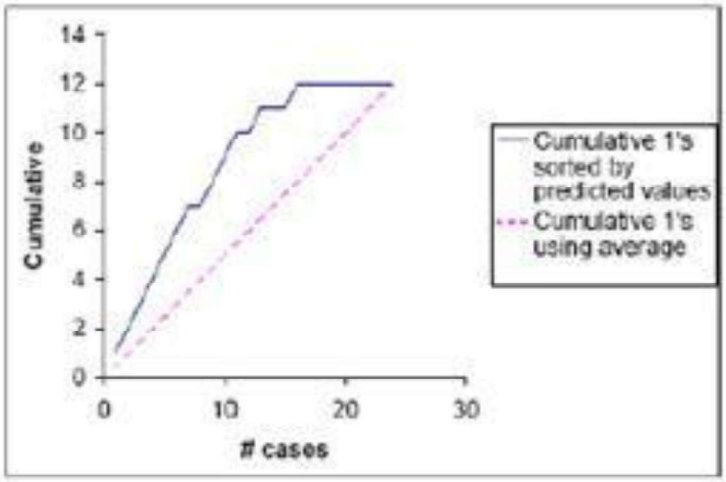

Figure 5.4. Li Curve

5.6.1. Limita ons of Li Chart

Predictable attribute in lift chart must be discrete. For this reason it can't be used to measure predictive performance for models which predicts continuous numeric values. To find out prediction line for any individual discrete value of predictable attribute, lift chart for that particular targeted value should be created. Multiple models can't be added to a lift chart if they do not share the same attributes.

5.7. Cumula ve Gain

It is evaluated 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.

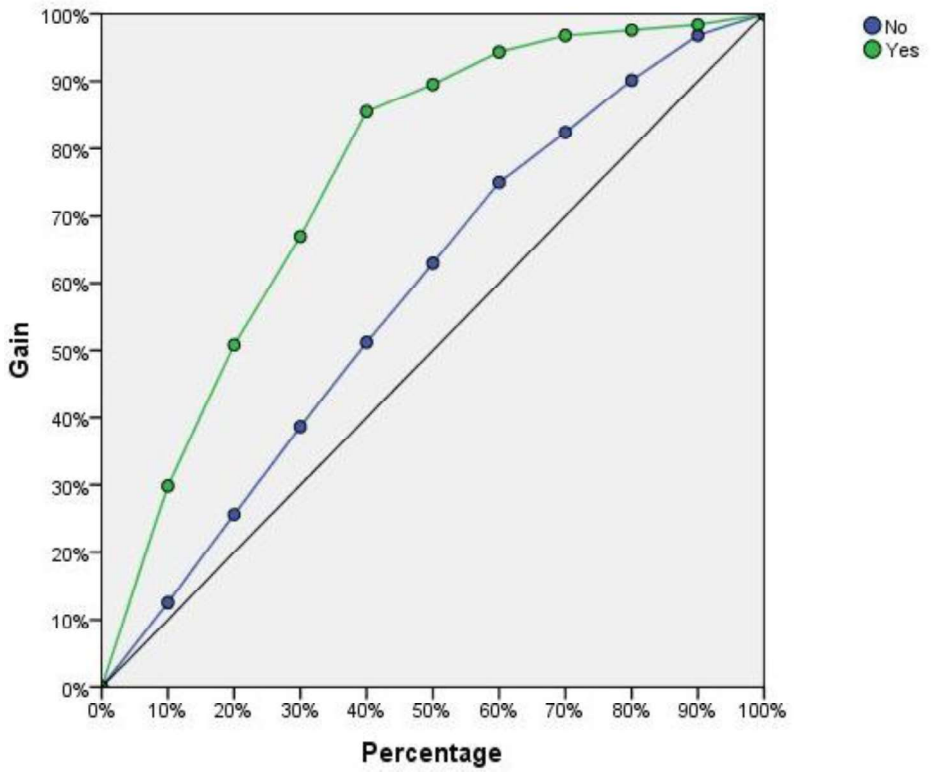

Figure 5.5. Cumula ve Gain Chart

It is the percentage of positive responses determined by the model across quantiles of the applied data. Cumulative gain for a given quantile is the ratio of the cumulative number of positive targets to the total number of positive targets.

Figure 5.5 shows the cumulative gain chart. The point in “yes” category (10%,30%) means that after scoring the data set if we arrange all the quantiles by predicted pseudo-probability of “yes” in descending order, one can expect the top 10% contain 30% of all the cases that take actually the “yes” category. Similarly the top 20% contain approximately 50% of the “yes” category. The diagonal line is known as base line. Farther above the base line, the gain will be greater.

5.7.1. Limitations of Cumulative Gain

Cumulative gain may fail to address the situation containing competing models where one model is not strictly above the other.

5.8. Probability Threshold

A continuous model can be transformed into a binary model by predicting only "Present" or "Absent" by introducing a threshold probability value such that any outcome greater than it is predicted to be "Present". True positive rates and false positive rates varies based on probability threshold values. Probability threshold for a quantile n is the minimum probability for the positive category to be included in this quantile or any quantiles preceding it. Cost threshold is reported if cost matrix is used, which is the maximum cost for the positive category to be included in the quantile or any quantiles preceding it.. This approach is not suitable in many circumstances notably when some records are not available [13]. Table 5.6 lists different methods that are used to select probability threshold.

Depending on the purpose of modeling the value of probability threshold is chosen. It is set to low value to identify larger area of potentially suitable habitat, if the purpose of modeling is to identify areas within which disturbance may impact a species negatively. Conversely, relatively high threshold is chosen, if the model was trying to identify potential introduction or reintroduction sites for an endangered species or species of recreational value. Choosing a high threshold reduces the risk of choosing unsuitable sites by identifying those areas with highest suitability [25].

5.8.1. Limitations of Probability Threshold

Probability threshold is used widely on uncertain database. Most of the literatures did not concentrate on the process of choosing the best threshold value. Too high or too low threshold value may cause either an empty result or too many results. Choosing the suitable threshold value is still a problem to a user.

Table 5.6. Some Published Methods for Se ng Threshold of Occurrence [14, 15]

Method	Procedure	Species Data Type
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 [16]
Fixed value	An arbitrary fixed value (e.g. probability = 0.5)	Presence only [17][18]
Sensitivity-Specificity equality	The threshold at which sensitivity and specificity are equal	Presence and Absence [16]
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 [19]
Maximize Kappa	The threshold at which Cohen's Kappa statistic is maximized	Presence and Absence [20][21]
Sensitivity-specificity sum maximization	The sum of sensitivity and specificity is maximized	Presence and Absence [22]
Lowest predicted value	The lowest predicted value corresponding with an observed occurrence of record	Presence only [23][24]
Average probability	The mean value across model output	Presence only [19]

5.9. Conclusion

This study presents the fact that any single approach as mentioned above cannot be used for all applications to judge them. Distribution of data, type of data, and type of model are the decisive factors to choose the parameters to predict the performance of predictive models. It is evident that in general, four measures, viz. precision, sensitivity, specificity, classification accuracy i.e. the entire confusion matrix should be recommended as benchmarking parameters to predict performance of an intelligent model without knowing the distribution of data. Area under receiver operating characteristic also plays an important role for predicting the performance of a model visually. Other parameters are used to select the best model by comparing a set of candidate models and in some cases to improve 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. Eisner R. 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. Swets J. A., "Measuring the accuracy of diagnostic systems", Science, Vol. 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.; Vol. 21, pp. 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, Vol. 20(1), pp. 37–46, 1960.
7. Landis J.R., Koch G.G., "The Measurement of Observer Agreement for Categorical Data", Biometrics, Vol. 33(1), pp 159-174, 1977.
8. Mc. Hugh M. L., "Inter-rater Reliability: the Kappa Statistic", Biochemia Medica, Vol – 22(3), pp. 276-282, 2012.
9. Akaike H., "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.
10. Akaike H., "A Bayesian Analysis of The Minimum AIC procedure", Ann. Inst. Statist. Math., Vol. 30(1), pp. 9-14, 1978.
11. G. Schwarz, "Estimating the Dimension of a Model, The Annals of Statistics", Vol. 6(2), pp. 461-464, 1978.
12. Jaffery T., Shirley X. Liu, "Measuring Campaign Performance by using Cumulative Gain Lift Chart", SAS Global Forum 2009, Paper 196, 2009.
13. Philips S. J. 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, Vol. 28, pp. 385-393, 2005.

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

16. Pearson R. G. et al., "Modeling Species Distributions in Britain: A Hierarchical Integration of Climate and Land-cover Data", Ecography, Vol. 27, 285-298, 2004.

17. Manel S. et al., "Comparing Discriminant Analysis, Neural Networks and logistic Regression for Prediction Species Distributions: A Case Study with a Himalayan River Bird", Ecological Modelling, Vol. 120, 337-347, 1999.

18. Robertson M. P. et al., "A PCA-based Modeling Technique for Predicting Environmental Suitability for Organisms from Presence Records", Diversity and Distributions, Vol. 7, 15-27, 2001.

19. Cramer J. S., "Logistic Models: From Economics and Other Fields", Cambridge University Press, 2003.

20. Huntley B. et al., "Modeling Present and Potential Future Ranges of Some European Higher Plants using Climate Response Surfaces", Journal of Biogeography, Vol. 22, pp.967-1001, 1995.

21. Elith J. et al., "Novel Methods Improve Prediction of Species' distributions from Occurrence Data", Ecography, Vol. 29, pp. 129-151, 2006.

22. Manel S. et al., "Evaluating Presences Absence Models in Ecology: The Need to Account for Prevalence", Journal of Applied Ecology, Vol. 38, pp.,921-931, 2001.

23. Pearson R. G. et al., "Model based Uncertainty in Species' Range Prediction", Journal of Biogeography, Vol. 33, pp. 1704-1711, 2006.

24. Phillips S. J. et al., "Maximum Entropy Modeling of Species Geographic Distributions", Ecological Modelling , Vol. 190, pp. 231-259, 2006.

25. Pearce J. et al., "Evaluating the Predictive Performance of Habitat Models developed using Logistic Regression", Ecological Modeling, Vol. 133, pp. 225 – 245, 2000.