Evaluation Measure Selection for Performance Estimation of Classifiers in Real Time Image Processing Applications

[1]Amandeep Singh, [2]Maninder Lal Singh
[1]Department of Electronics Technology, Guru Nanak Dev University Amritsar, India
[2]Department of Electronics Technology, Guru Nanak Dev University Amritsar, India
Email: [1]amandeep.singh.sodha@gmail.com, [2]mlsingh7@gmail.com

Abstract: Deciding the criterion for the performance evaluation of a classifier plays a vital role in the selection procedure of a classifier for a certain problem. These criteria empower the researchers to do the selection of a classifiers for effective classifications of unseen data from a range of classifying algorithms. A great number of different measures are currently available for the classification problems based on binary, flat or undistributed data such as in case of images. However in case of hierarchical classifications, where the number of classes to be identified are more than two, the evaluation of a classifier becomes more and more intricate as the classes to be differentiated, are hierarchically attached. The topic of focus of this paper is to provide a knowledge flow which a researcher can use while dealing with such real time based problems where the accuracy and efficiency of a classifier are the major concerns. The problem of interest while discussing the different aspects of various evaluation measures, was the color prediction of paddy crop plant leaf for its health characterization.

Keywords: Machine Learning; Confusion Matrix; ROC; AUC; Cost Curve; Accuracy

1. Introduction

In the epoch of technology, machine learning and hence machine vision has changed the way of living beings in almost every aspect. Making a very perfect replacement of humans in all areas of applications, it has established the new benchmarks. Although a number of these algorithms has touched the new heights but still there are some points where the machine learning is lacking than the neural minds. For instance, let the example of undistributed and unseen data. When some algorithm has to be designed to work on such data based applications, where some sort of classification is needed, clearly it becomes a difficult task for the designer to achieve high accuracy rates. On the other side, everyone is familiar with the classifiers used for such classifications. While training and testing of these classifiers it becomes necessary to evaluate the performance of a particular classifier so that one can use it for achieving the goals of research. A large number of evaluating measures are available which work on different characteristics of a classifier. Thus it is the on-going topic of research to select such measures, even for binary class classifiers[1]. While handling the problem of multiclass classifications, where data which has to be processed, is from hierarchically related classes, the measures used for the binary classifications are used, which in turn does not characterize the classifier with required accuracy and thus become inadequate. The motivation behind this paper is to resolve this problem by providing a specific detail about different measures which can be used for the classifier performance evaluation, used in multiclass classifications. The underhand problem is a real time case, chosen from daily life, in which measures for the classifiers used in prediction of color of paddy plant leaf are discussed. Amandeep Singh and Maninder Lal Singh [2], working in same classification area where prediction of color is done in reference with the Leaf Color Chart (LCC), there is need of such measures as the data has to be classified in at least six classes (Different shades of green).

2. DEFINING CLASSIFIER EVALUATION

Analysing the history, it becomes crystal clear that a number of researchers are working on the said topic. However there is an ambiguity which arises from the definition of classifier performance and hence changes the evaluation process. Generally it is said that a best classifier is the one which accurately classifies the data. But it should be defined as that best one is the one which actually has best generalizing capability of test data, on the basis of training data. Here it should be noticed that training dataset is the data on basis of which a classifier maps/matches the features of test data onto the features of different classes. The training data should have following features:

1. Differentiable corresponding to different classes

2. Finite but fuzzy values
3. A set of sufficient (Not too many or too less) parameters
4. Should be extracted randomly from a large number of population

All above mentioned points not only train a classifier but also boost up its generalizing capability. If the training data has above specifications then it will also ensure least danger of over fitting which is result of training data being too precise. Following diagrams show the block diagram of classifier evaluation process and results corresponding to different types of training data:

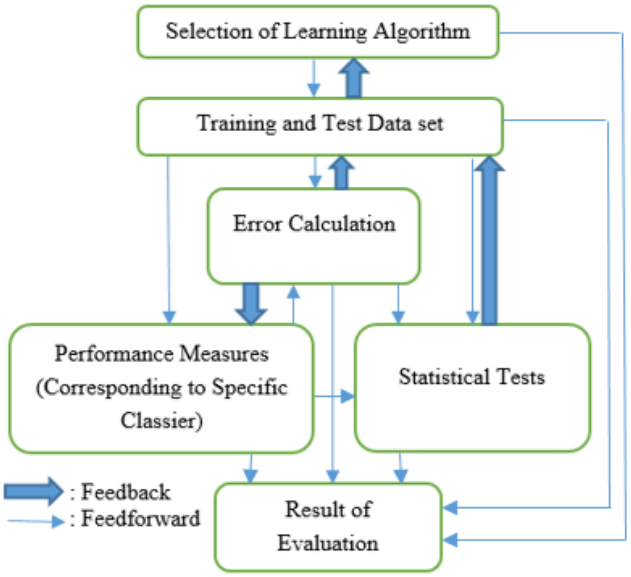

Fig. 1. Evaluation Process

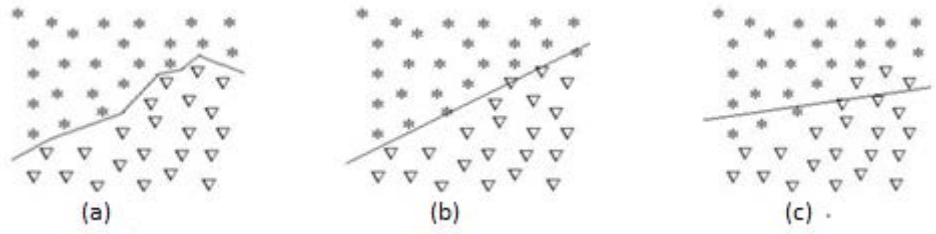

Fig. 2. a. Over fit Training Data for ClassifierFig. b. Fit Training Data for ClassifierFig. c. Under fit Training Data for Classifier

Dealing with the problem of classification of leaves of different shades, for implementing the first step of evaluation process, the training parameters were calculated. These parameters include:

1. R, G, B content of Image
2. Hue value of Image
3. Standard Deviation
4. Variance of Image
5. Mean Square Error

These parameters were found to be having different values for different classes using which a training dataset set can be generated. Here it has to be noticed that only the data which can be used as backbone of the class prediction algorithm is Red, Green and Blue content and while dealing with the human perception is hue content. Other parameters here are added make it a fit training data for classifier. If only R,G,B and Hue values are used then the training data becomes under fit which in turn decreases the generalization capability of the classifier. In other words using only four parameters for the

training, result in 100% accuracy with all classifiers which is practically not possible to obtain, adding to the conclusion that some more parameters has to be added as the training data is lacking in generalization values.
After getting the training data set, now the second step was to check the error rate of a classifier. Here is step where performance measures were selected after going through a lot of literature survey for multiclass classifications. A study of hierarchical evaluation measures has been proposed by Kiritchenko in 2005, with a distinct accent [3]. The bulk of classification problems in the literature comprises flat classification, where each example is assigned to a class out of a finite set of flat classes. However, there are more multifaceted classification problems, in which include classes to be predicted are hierarchically related [4, 5]. Also a number of flat measures are available and are proposed by researchers of university of Waikato [6] who have also designed a data mining tool named as Weka which also evaluates the classifiers used in binary and multiclass problems. Another conjoint evaluation measure used in binary classification problems is the ROC (Receiver Operating Characteristics), which narrates sensitivity and specificity. Although ROC curves were formerly developed for two-class problems, they have also been generalized for multi-class problems [4]. Other than these depth dependent measures [7], Semantic based measures [8], Hierarchy based Measures [9] have beautifully summarized different evaluation measures for binary class and multiclass classifiers.

3. PRFORMANCE MEASURES

At last from above all survey a summary of different measures have been obtained and is given in figure 3 [10]. For the multiclass classification problem of paddy leaves classification, the different measures which are proposed are discussed thereafter.

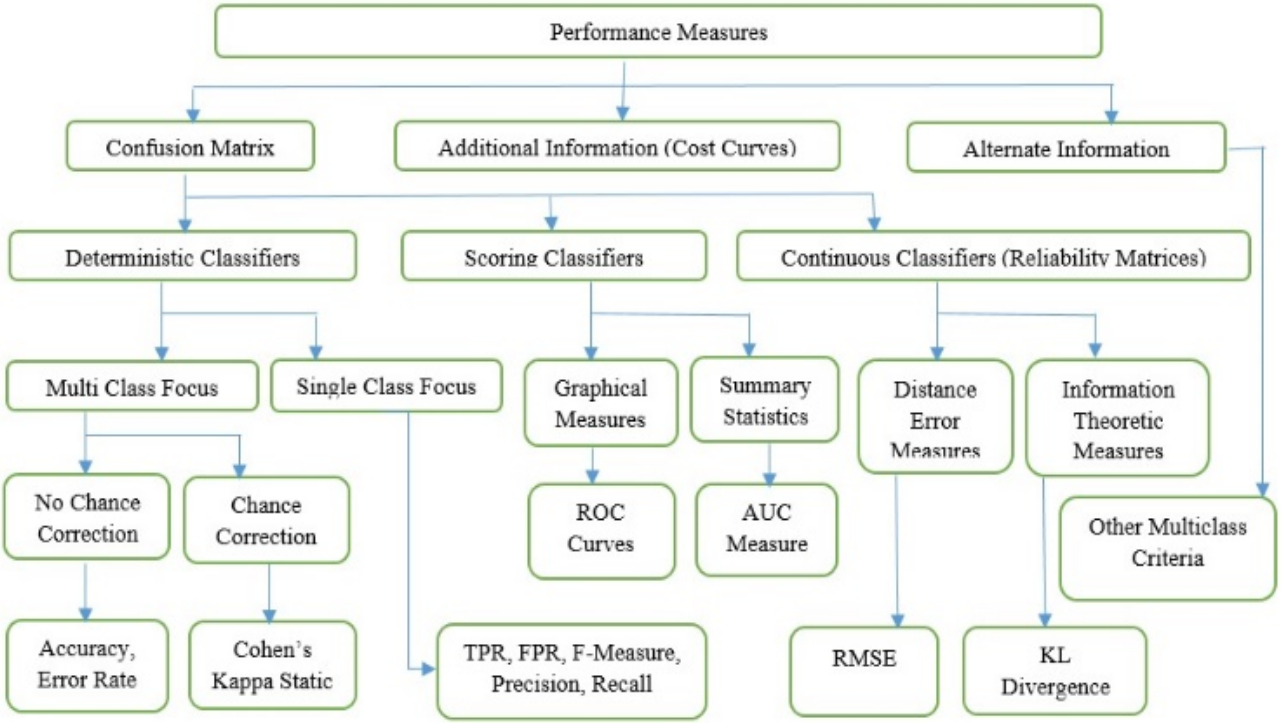

Fig. 3. Evaluation Measures

A. Confusion Matrix: It deals with a number of performance measures which further are helpful in classifier evaluation. Sometimes also known as contingency table or accumulation matrix, the confusion matrix is an excellent tool which can be used in binary class as well as multiclass problems. In other words it can also be defined as the tool that consents conception of the performance of a process, typically based on supervised learning. Figure 4 shows a confusion matrix for with its performance measuring terms. The variables a,b,c and d will be further used for the calculation of some another performance evaluation parameters which are dependent on confusion matrix.

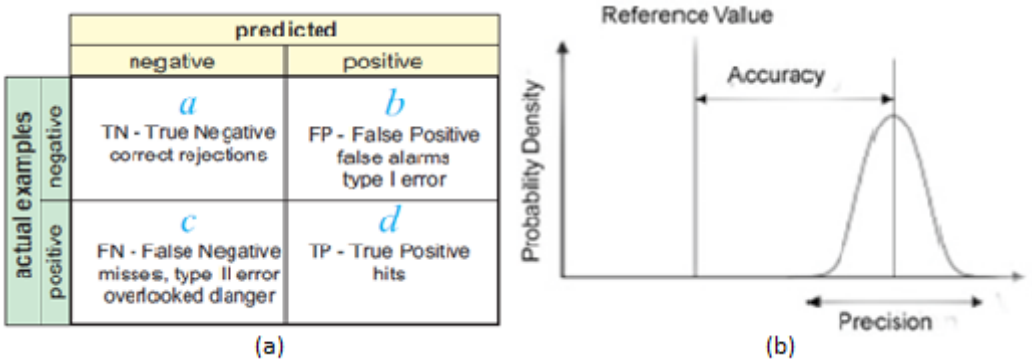

Fig. 4. a. Confusion Matrix b. Dictation of Accuracy and Precision

B. Accuracy: Overall efficacy of a classifier is defined as the accuracy. Sometimes also known as the level of measurement without any essential constraint. It is a measure that yields true and consistent results. It can be calculated from confusion matrix as:

$$Accuracy=(a+d)/(a+b+c+d)$$

C. Precision: The measurement that provides consistent results when done repeatedly is called precision. Sometimes it is confused with accuracy but the figure 6 clarifies the difference. Again from the Confusion matrix the precision can be estimated as:

$$Precision= d/(b+d)$$

D. True Positive Rate: This parameter is a statistical measure of performance of a binary class classifying algorithm. It provides an idea about proportion of positive detections that are correctly identified. A synonym for it is sensitivity and in certain area of statistical calculations also known as recall. It can be predicted from the confusion matrix as:

$$TPR= d/(c+d)$$

E. Specificity: The term which gives the proportion of negatives that are correctly found by some algorithm. Confusion matrix helps again in calculation of the parameter as:

$$Specificity= a/(a+b)$$

F. F-Measure: It is a measure which is used for the measuring accuracy of a classifier taking into account the values of recall and presision. It can be interpreted as harmonic mean of precision and recall. Not directly nut indirectly it is again can be computed from confusion matrix. However it is calculated as:

$$F\text{-}Measure=2*Precision*Recall/(Precision+Recall)$$

G. Problem with Scalar Measures: From the above all measures it can be concluded very clearly that don't provide enough information such as how the errors are distributed across all the classes, how the behaviour of a classifiers varies under different testing environments etc. However the measures as True Positive Rate and False Positive Rate etc. discussed in previous sections are much more informative than the scalar ones. But the difficulty to study two measures simultaneous makes them little bit typical while handling.

H. Receiver Operating Characteristics (ROC): A better solution to the problem mentioned in above section has been resolved by Receiver Operating Characteristics (ROC) which informs about:

a) Performance for all possible wrong classification costs.
b) Performance for all possible class ratios.
c) All conditions under which one particular classifier is better than the other one.

Some characteristics of ROC curve includes their usefulness in evaluation of dichotomic classifiers, characterising capability to show degree of overlapping classes for a single feature and provision of providing decision on the basis of single threshold values. Figure 5(a) explains the ROC space in detail and how to drive decisions based on ROC curves.

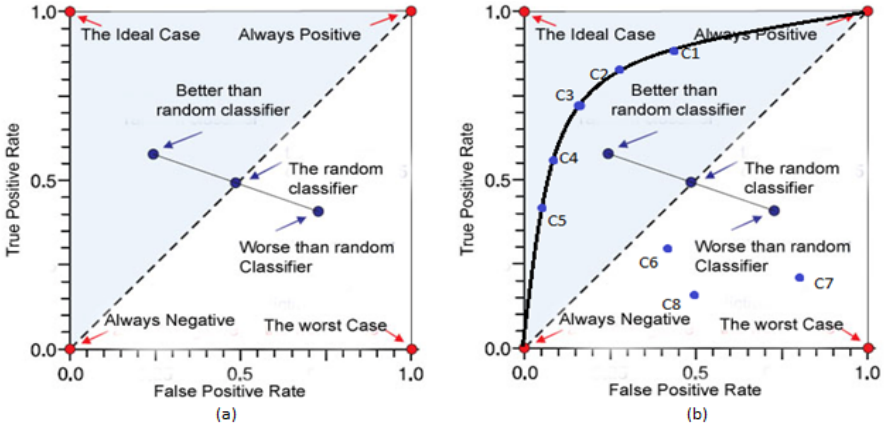

Fig. 5. a. ROC Curve Dictation b. ROC Curve Based Decisions

Now if we plot the performances of actual classifiers then the better ones will make a curve with the shape of convex hull. All the classifiers with this shape can be treated as the better ones than those which are under the convex hull. Diagram 5(b) depicts the case where classifiers from C1 to C5 can be declared as the best ones than the classifiers from C6 to C8.
The main advantage of a ROC curve is that enables the user to distinguish between discriminability and verdict unfairness. If an assumption of Gaussian distribution is taken into account then another term named as Bayesian Error Rate for a classifier can also be calculated, whose explanation is beyond the limits of this paper. However ROC curve is insensitive to problem based on skew class distribution and misclassification cost. Some researchers now a days has adopted the AUC (Area under ROC curve) which makes it easy to make decision.
If A and B are two classifiers such as:

$$AUC(A) > AUC(B)$$

Then it can be concluded that classifier A is better than the other one.

I. Measure for Deterministic, Scoring and Probabilistic Classifiers: In some cases it has been noticed that correct classification could be a consequence of accidental concordance between the classifier's output and the label generation process. To avoid this ambiguity another measure has been proposed by Cohen discussed in later section.

J. Cohen's Kappa Static: It is a robust measure than calculates percentage of agreement, as it deals with the agreement occurring by chance. From mathematical point of view it estimates the agreement among two raters who each classify X number of items into C mutually exclusive classes. It can be calculated as:

$$\varkappa = (P0 - PeC)/(1 - PeC)$$

where P0 represents the probability of overall agreement between the classifier and the true process and PeC represents the chance agreement.

Cost Curves are the graphical measures which deal with the scoring classifiers. These are more practical then ROC curves and AUCs as they estimate the probabilities of a particular class for which one classifier is preferable over the other. An example of a cost curve has been shown in figure 6 for better understanding of concept.

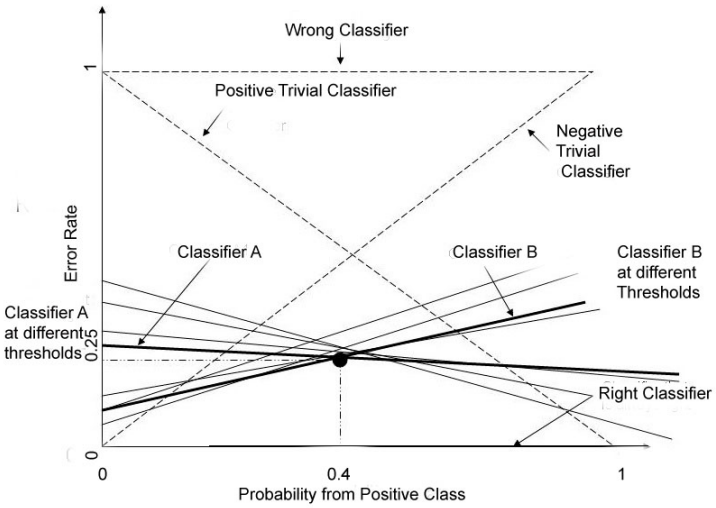

Fig. 6. Cost Curve

K. Root-Mean Squared Error (RMSE):In case of probabilistic classifiers the measure used is Root-Mean Squared Error (RMSE). It is usually used for regression. The formula for the RMSE is:

$$\mathrm{RMSE(f)} = \mathrm{sqrt}\left(\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2\right)$$

Where m is the number of test examples, f(xi), the classifier's Probabilistic output on xi and yi the actual label [REFERENCE]. It is a measure which is based on distance calculations between the actual class and the identified class.

4. CONCLUSION AND FUTURE SCOPE

As the topic of concern as defined in the objectives of this paper was to select evaluation measures for the classifiers used to classify the color of paddy plant leaf for the health analysis. So the measures for the task were supposed to be related to multiclass classifications. But here as the color prediction algorithm has been trained to do so on certain protocols which include processing of the images in binary domain. So because of this reason the evaluation criteria has been chosen in such a way that it may evaluate the performance in binary class cases as well as the multi class cases. So following measures has been finalized to get the objectives achieved:

1. TPR (True Positive Rate) and FPR (False Positive Rate)
2. Accuracy
3. Precision
4. Recall
5. F-Measure
6. ROC (Receiver Operating Characteristics)
7. AUC (Area Under ROC Curve)
8. Kappa Static
9. MAE (Mean Square Error)
10. RMSE (Root Mean Square Error)

Summarizing about future research, using these measures the performance of a number of classifiers available under different classes like Bayes, Functions, Lazy, Meta, Rules, Trees etc. would be evaluated for the classifications of leaf color in reference with leaf color chart having six shades (six classes to be identified by classifier) of green using tools like Weka 3.6 and MATLAB.

5. Acknowledgment

This work is supported by Electronics and Communication Department of Guru Nanak Dev University, Amritsar, Punjab, by providing excellent laboratories (Computer Lab) and MATLAB software for the research work.

References

1. Sokolova, M.; Japkowicz, N.; and Szpakowicz, S. Beyond, "Accuracy, F-score and ROC: A Family of Discriminant Measures for Performance Evaluation",in Proceedings of the 'AAAI'06 workshop on Evaluation Methods for Machine Learning', 24–29, 2006
2. Amandeep Singh, Maninder Lal Singh, "Automated Color Prediction of Paddy Crop Leaf using Image Processing", in IEEE Xplore via 'International IEEE Conference TIAR 2015, ISBN: 978-1-4799-7758-1', 2013
3. Kiritchenko, S. , "Hierarchical Text Categorization and Its Application to Bioinformatics", At'Ph.D. Dissertation School of Information Technology and Engineering, Faculty of Engineering', University of Ottawa, Ottawa, Canada, 2005
4. Everson, R. M., and Fieldsend, J. E. , "Multi-class ROC Analysis from a Multi-Objective Optimisation Perspective" in'Pattern Recogn. Lett. 27(8):918–927, 2006.
5. Sun, A.; Lim, E. P.; and Ng,W. K., "Hierarchical Text Classification Methods and Their Specification. Cooperative Internet Computing" 2003.
6. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, 'The WEKA Data Mining Software: An Update' SIGKDD Explorations, Volume 11, Issue 1, 2009.
7. Blockeel, H.; Bruynooghe, M.; Dzeroski, S.; Ramon, J.; and Struyf, J.,"Hierarchical Multi-Class Classification" in 'Proceedings of the ACM SIGKDD Workshop on Multi-Relational Data Mining (MRDM)', PP21–35, 2002.
8. Sun, A., and Lim, E. P., "Hierarchical Text Classification and Evaluation", In 'Proceedings of the 2001 IEEE International Conference on Data Mining, PP521–528. IEEE Computer Society' Washington, DC, USA, 2001
9. Ipeirotis, P.; Gravano, L.; and Sahami, M., "Probe, Count, and Classify: Categorizing Hidden Web Databases", In 'Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data',PP 67–78. ACM Press New York, NY, USA, 2001
10. Nathalie Japkowicz & Mohak Shah, "Evaluating Learning Algorithms: A Classification Perspective", Cambridge University Press, 2011
11. Eduardo P. Costa, Ana C. Lorena Andr´e C. P. L. F. Carvalho Alex A. Freitas, "Association for the Advancement of Artificial Intelligence" 2001.
12. Jose A. Lozano, Guzman Santafe, Inaki Inza,"Intelligent Systems Group " at 'International Conference on Machine Learning and Applications (ICMLA 2010)' The University of the Basque Country December 12-14, 2010
13. Bishop, C.M., " Pattern Recognition and Machine Learning (Information Science and Statistics)", in Springer Verlag New York, Inc., Secaucus, NJ, USA (2006)
14. Brodersen, K.H., Mathys, C., Chumbley, J.R., Daunizeau, J., Ong, C.S., Buhmann, J.M., Stephan, K.E., " Bayesian Mixed-Effects Inference on Classification Performance in Hierarchical Data Sets", in'Journal of Machine Learning Research 13 (Nov. 2012)'PP3133 – 3176. 2012
15. Carrillo, H.: GBAC. In: http://www.mloss.org/software/view/447/. (2013)
16. Marina Sokolova, Marina Sokolova, " A Systematic Analysis of Performance Measures for Classification Tasks", in'Information Processing and Management'PP427–437, 2009
17. Goutte, C., & Gaussier, E. , "A probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation", In 'Proceedings of 27th European Conference on IR Research' PP 345–359, 2005.
18. Lachiche, N., & Flach, P. A.,"Improving Accuracy and Cost of Two-Class and Multi-Class Probabilistic Classifiers using ROC Curves", In 'Proceedings of ICML'2003', PP 416–423, 2003
19. Li, T., Zhang, C., &Zhu, S., "Empirical Studies on Multi-Label Classification", In 'Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence', PP 86–92, 2006.
20. hu, S., Ji, X., Xu, W., & Gong, Y., " Multi-Labelled Classification using Maximum Entropy Method", In 'Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval, PP 274–281, 2005.