www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 7 July, 2014 Page No. 7217-7220

A Review on Evaluation Measures for Data Mining Tasks

B. Kiranmai[1], Dr. A. Damodaram[2]

[1] Assoc. Professor, CSE Dept.,
Nishitha College of Engineering &Technology
Greater Hyderabad, India
kiranmaimtech@gmail.com

[2] Professor in CSE, Director AAC,
JNTUH, Kukatpally
Hyderabad
damodaramma@rediffmail.com

Abstract: ***The cutting edge of technology is data mining. We apply data mining techniques in various fields and the results are enormous. Assessment is important to check your quality, and these metrics or measures play a vital role in guiding the work. This paper summarizes various evaluation measures criteria for assessment of data mining tasks. In this paper data mining functionalities (classification, clustering, Association rule mining) assessment metrics are discussed.***

Keywords: data mining; classification; clustering; Association; evaluation;

1. Introduction

We use data mining techniques for financial analysis, retail Industry, biological data analysis, to find fraud detection, intrusion detection, text mining, and other scientific applications. Evaluating the performance of a data mining technique is a fundamental aspect of machine learning. Evaluation method is the yard stick to examine the efficiency and performance of any model. The evaluation is important for understanding the quality of the model or technique for refining parameters in the iterative process of learning and for selecting the most acceptable model or technique from a given set of models or techniques. There are certain criteria for evaluating models for different tasks. The most widely used different data mining techniques are Classification, Clustering ,Association.

2. Evaluation Measures of a Classification Model

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data.[2]

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The associated class label of each test tuple is compared with the learned classifier's class prediction for that tuple. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known.[2]

2.1 Cross Validation

In ***k***-fold cross-validation, the initial data are randomly partitioned into *k* mutually exclusive subsets or "folds," *D*1, *D*2, : : : , *Dk*, each of approximately equal size. Training and testing is performed *k* times. In iteration *i*, partition *Di* is reserved as the test set, and the remaining partitions are collectively used to train the model. That is, in the first iteration, subsets *D*2, : : : , *Dk* collectively serve as the training set in order to obtain a first model, which is tested on *D*1; the second iteration is trained on subsets *D*1, *D*3, : : : , *Dk* and tested on *D*2; and so on. Unlike the holdout and random subsampling methods above, here, each sample is used the same number of times for training and once for testing. For classification, the accuracy estimate is the overall number of correct classifications from the *k* iterations, divided by the total number of tuples in the initial data.[1]

Leave-one-out is a special case of *k*-fold cross-validation where *k* is set to the number of initial tuples. That is, only one sample is "left out" at a time for the test set. In stratified cross-validation, the folds are stratified so that the class distribution of the tuples in each fold is approximately the same as that in the initial data.[1]

2.2 Sensitivity and Specificity

Sensitivity is also referred to as the *true positive (recognition) rate* (that is, the proportion of

positive tuples that are correctly identified), while specificity is the *true negative rate* (that is, the proportion of negative tuples that are correctly identified).

$$\text{Sensitivity} = \frac{t\_pos}{pos} \quad (1)$$

$$\text{Specificity} = \frac{t\_neg}{neg} \quad (2)$$

t_pos refers to number of true positives that are correctly classified

pos is the number of positive tuples.

t_neg refers to number of true negatives that were correctly classified

neg is the total number of negatives.

2.2.1 Precision

Precision is fraction of retrieved instances that are more relevant than irrelevant.[4]

In a classification task, the precision for a class is the number of **true positives** (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and **false positives**, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and **false negatives**, which are items which were not labeled as belonging to the positive class but should have been).[4]

$$\text{Precision} = \frac{t\_pos}{(t\_pos + f\_pos)} \quad (3)$$

t_pos is true positive,

f_pos is false positive.

2.2.2 Recall is the fraction of retrieved instances that are retrieved.[4]

$$\text{Recall} = \frac{t\_pos}{(t_{neg} + f\_neg)} \quad (4)$$

t_neg is true negatives

f_neg is false negatives

Accuracy is defined as a function of sensitivity and specificity

$$\text{Sensitivity} \frac{pos}{(pos+neg)} + \text{specificity} \frac{neg}{(neg+pos)} \quad (5)$$

2.3 Confusion Matrix

The *confusion matrix* is a useful tool for analyzing how wel l your classifier can recognize tuples of different classes.
Given *m* classes, a confusion matrix is a table of at least size *m* by *m*. An entry, *CMi*, *j* in the first *m* rows and *m* columns indicates the number of tuples of class *i* that were labeled by the classifier as class *j*. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, from entry *CM*1, 1 to entry *CMm*, *m*, with the rest of the entries being close to zero. The table may have additional rows or columns to provide totals or recognition rates per class.

2.4 Hold out and Random sub sampling

In hold out method, the given data are randomly partitioned into two independent sets, a *training set* and a *test set*. Typically, two-thirds of the data are allocated to the training set, and the remaining one-third is allocated to the test set. The training set is used to derive the model, whose accuracy is estimated with the test set.[2]

Random subsampling is a variation of the holdout method in which the holdout method is repeated *k* times. The overall accuracy estimate is taken as the average of the accuracies obtained from each iteration.[2]

2.5 Bootstrap

Unlike other methods discussed above, in which the sampling is done without replacement. In Bootstrap method the training records are sampled with replacement i.e. a record already chosen for training is put back into original pool of records so that it is equally likely to be redrawn.[2]

2.6 Bagging

Bagging is nothing but Bootstrap Aggregation. The bagging algorithm creates an ensemble of models (classifiers or predictors) for a learning scheme where each model gives an equally-weighted prediction.[2]

Given a set D, of *d* tuples, bagging works as follows. For iteration *i* (*i* = 1, 2, : : : , *k*), a training set,*Di*, of *d* tuples is sampled with replacement from the original set of tuples ,*D*. The bagged classifier often has significantly greater accuracy than a single classifier derived from *D*, the original training data.

2.7 ROC Curves

ROC curves are a useful visual tool for comparing two classification models. The name ROC stands for *Receiver Operating Characteristic*. ROC curves are two dimensional graphs that visually depict the performance and performance trade off of a classification model. In order to construct ROC curve two performance metrics are to be used. They are true positive rate (TPR) and false positive rate (FPR)

$$\text{True Positive Rate} = \frac{t\_pos}{(t-pos+f_{neg})} \quad (6)$$

$$\text{False positive Rate} = \frac{f\_pos}{(f_{pos}+t_{pos})} \quad (7)$$

ROC graphs are constructed by plotting the true positive rate against the false positive rate.
An ROC curve for *M* is plotted as follows. Starting at the bottom left-hand corner (where the true positive rate and false-positive rate are both 0), we check the actual class label of the tuple at the top of the list. If we have a true positive (that is, a positive tuple that was correctly classified), then on the ROC curve, we move up and plot a point. If, instead, the tuple really belongs to the 'no' class, we have a false positive. On the ROC curve, we move right and plot a point. This

process is repeated for each of the test tuples, each time moving up on the curve for a true positive or toward the right for a false positive.[1]

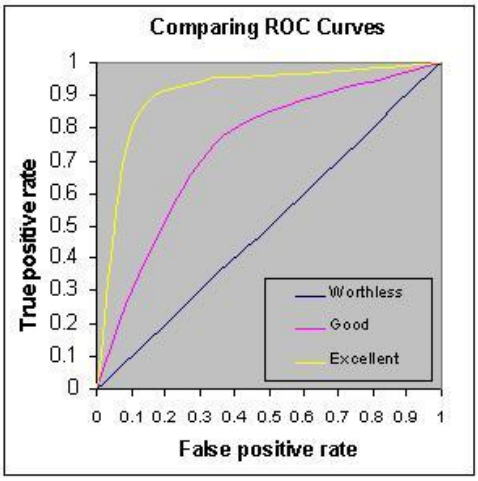

Fig 1 Comparing ROC curve[5]

3. Evaluation of a Cluster Model

Clustering techniques consider data tuples as objects. They partition the objects into groups or *clusters*, so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters. Similarity is commonly defined in terms of how "close" the objects are in space, based on a distance function.[7]

Clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of *maximizing the intra class similarity and minimizing the interclass similarity*. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.[8]

The quality of a cluster is represented by its diameter, the maximum distance between any two objects in a cluster.

Centroid distance is an alternative measure of cluster quality and is defined as average distance of each cluster object from the cluster centroid.

4. Evaluation methods for Association Rule Mining

Frequent item set mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes, such as catalog design, cross-marketing, and customer shopping behavior analysis.

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used.

The two thresholds are minimum support and minimum confidence were originally proposed by [Agrawal and Srikanth 1994].[3]

3.1 Minimum Support

The support supp(x) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.

3.2 Minimum Confidence

Confidence can be interpreted as an estimate of the probability P(Y / X), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

The confidence of a rule is defined as

$$\text{Conf}(X \rightarrow Y) = Supp(X \cup Y)/Supp(X) \qquad (8)$$

3.3 Lift

The lift of a rule is defined as

$$\text{Lift}(X \rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(Y) * Supp(X)} \qquad (9)$$

Ratio of the observed support to that expected if X and Y were independent.

3.4 Conviction

The conviction of a rule is defined as

$$\text{Conv}(X \rightarrow Y) = \frac{1-supp(X)}{1-conf(X \rightarrow Y)} \qquad (10)$$

Ratio of expected frequency that X occurs without Y(that is to say the frequency the rule makes an incorrect prediction) if X and Y were independent divided by the incorrect predictions .[1]

Conclusion

Section II, III, IV summarized various measures of Classification, Clustering, and Association rule mining techniques. Choosing a metric for evaluation depends on the type of data, and application. But in practice to achieve accuracy use combination of one or more measures. Measuring assessment of classification model, the popularly used metrics are precision and recall. Bagging and boosting are ensemble methods with which accuracy can be increased. For comparison two or more models, ROC curve are appropriate. For Clustering, measuring quality mainly depends on distance between each object. Minimum support and minimum confidence are the metrics used for association rule mining.

Acknowledgement

Thanks to my Director Sir Prof Vijay Kumar, Nishitha College of Engineering &Technology for his encouragement and support. Thanks to colleagues of CSE Dept.

References

[1] http://iasri.res.in/ebook/win_school_aa/notes/Evaluation_Measures.pdf

[2] Jiawei Han and Micheline Kamber Data Mining Concepts and Techniques Second Edition Morgan Kauffman Publishers ,2006

[3] Liqiang Geng and Howard J.Milton Interstingness Measures For Data Mining ACM Computing Surveys, Vol 38, No. 3, Article 9,Publication date :September 2006

[4] http://en.wikepedia.org/wiki/Precision_and recall

[5] http://google.co.in/ROCcurve

[6] VAILLANT, B., LENCA, P., AND LALLICH, S. 2004. A clustering of interestingness measures. In *Proceedings of the 7th International Conference on Discovery Science* (*DS 2004*). Padova, Italy. 290–297.

[7] http://en.wikepedia.org/wiki/datamining techniques

[8] BAYARDO, R. J. AND AGRAWAL R. 1999. Mining the most interesting rules. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining* (*KDD-99*). San Diego, CA. 145–154.

[9] HILDERMAN, R. J. AND HAMILTON, H. J. 2001. *Knowledge Discovery and Measures of Interest*. Kluwer Academic, Boston, MA.