Classification and Prediction

——Classification Accuracy——

徐华

清华大学 计算机系 智能技术与系统国家重点实验室

xuhua@tsinghua.edu.cn

1

Classification and Prediction

- Basic Concepts
- Issues Regarding Classification and Prediction
- Decision Tree
- Bayesian Classification
- Neural Networks
- Support Vector Machine
- Support Vector Machine
- K-Nearest Neighbor
- Associative classification
- Classification Accuracy

2

Classification Accuracy: Estimating Error Rates

- **Related to one special test data set.**
- **Accuracy(准确率), Precision (精确率) and Recall (召回率)**
  - The corresponding computation formula are listed as the following.
  - Accuracy （准确率，针对所有类别而言，平均分类效果）
  - Precision （精确率，针对某个类别而言）
  - Recall （召回率，针对某个类别而言）
- **F1 Score**

(精确率与召回率的调和平均)

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \text{(方便理解)} \quad ------(1)$$

$$F1 = \frac{2Precision * Recall}{Precision + Recall} \text{(标准公式)} \quad ----(2)$$

真实值 \ 预测值	正例	负例
正例	TP (真正例)	FN (假负例)
负例	FP (假正例)	TN (真负例)

准确率: $Acc = \frac{TP+TN}{TP+FN+TN+FP}$

召回率: $Recall = \frac{TP}{TP+FN}$

精确率: $Precision = \frac{TP}{TP+FP}$

真实值 \ 预测值	类别1	类别2	类别3
类别1	R_{11}	R_{12}	R_{13}
类别2	R_{21}	R_{22}	R_{23}
类别3	R_{31}	R_{32}	R_{33}

R_{ij}: 表示真实值为类别i，预测值为类别j的样本数量

准确率: $Acc = \frac{\sum_{i=1}^{3} R_{ii}}{\sum_{i=1}^{3}\sum_{j=1}^{3} R_{ij}}$

类别i的召回率: $Recall(i) = \frac{R_{ii}}{\sum_{j=1}^{3} R_{ij}}$

类别i的精确率: $\text{Precision}(i) = \frac{R_{ii}}{\sum_{j=1}^{3} R_{ji}}$

3

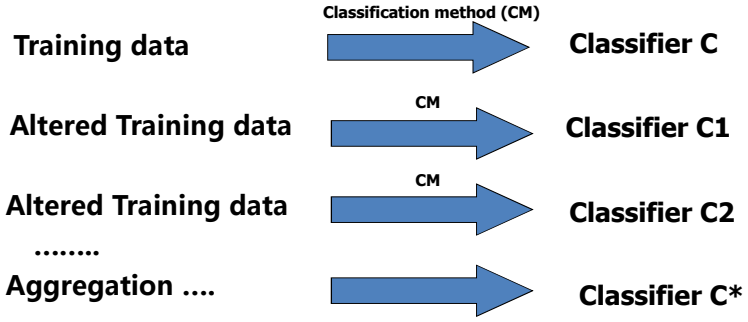
Classification Accuracy: Estimating Error Rates

- **Partition: Training-and-testing**
  - **use two independent data sets, e.g., training set (2/3), test set(1/3)**
  - **used for data set with large number of samples**
- **Cross-validation**
  - **divide the data set into *k* subsamples**
  - **use *k-1* subsamples as training data and one sub-sample as test data—*k*-fold cross-validation**
  - **for data set with moderate size**

4

Bagging and Boosting

- **General idea :** sampling



5

Bagging

- Given a set S of s samples
- Generate a bootstrap(引导程序) sample T from S. Cases in S may not appear in T or may appear more than once.
- Repeat this sampling procedure, getting a sequence of k independent training sets
- A corresponding sequence of classifiers C1,C2,…,Ck is constructed for each of these training sets, by using the same classification algorithm
- To classify an unknown sample X, let each classifier predict or vote
- The Bagged Classifier C* counts the votes and assigns X to the class with the "most" votes

6

Boosting Technique — Algorithm

- **Assign every example an equal weight *1/N***
- ***For t = 1, 2, …, T Do***
  - Obtain a hypothesis (classifier) $h^{(t)}$ under $w^{(t)}$
  - Calculate the error of *h(t)* and re-weight the examples based on the error . Each classifier is dependent on the previous ones. Samples that are incorrectly predicted are weighted more heavily
  - Normalize $w^{(t+1)}$ to sum to 1 (weights assigned to different classifiers sum to 1)
- **Output a weighted sum of all the hypothesis, with each hypothesis weighted according to its accuracy on the training set**

7

Boosting Technique — Algorithm

- k, the number of classifiers
- Di, the ith training set sampling from S,
- Mi, the classifier of ith model corresponding to Di
- n, the number of the samples in each Di

Boosting Algorithm:

```
1  for i=1 to k
2      construct Di;
3      initialize weights of each sample(1/n);
4       construct Mi using Di;
5       evaluate classifier err(Mi); (if err(Mi)>threshold, then go to 2)
6       update the weight of each sample;
7  endfor
8  calculate the weight for each classifier Mi
```

8

Boosting Technique — Algorithm

- construct Di
  - Sampling with the replacement strategy
- Evaluate err(Mi)

$$err(M_i) = \sum_{i=1}^{d} w_j err(x_j) \qquad err(x_j) = \begin{cases} 1 & misclassification\ for\ x_j \\ 0 & rightclassification\ for\ x_j \end{cases}$$

- Update weights of sample for x_j

$$w_j \times err(M_i)/(1 - err(M_i))$$

- Calculate the weights of k classifiers

$$\log \frac{(1 - err(M_i))}{err(M_i)}$$

9

Summary

- Classification is an extensively studied problem (mainly in statistics, machine learning & neural networks)
- Classification is probably one of the most widely used data mining techniques with a lot of extensions
- Scalability is still an important issue for database applications: thus combining classification with database techniques should be a promising topic
- Research directions: classification of non-relational data, e.g., text, spatial, multimedia, etc..

10

Thanks !