Classification and Prediction

——Issues Regarding Classification and Prediction——

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
- K-Nearest Neighbor
- Associative Classification
- Classification Accuracy

2

Issue 1: Data Preparation

- Data cleaning
  - Preprocess data in order to reduce noise and handle missing values
- Relevance analysis (feature selection)
  - Remove the irrelevant or redundant attributes
- Data transformation
  - Generalize and/or normalize data

Issue 2: Evaluating Classification Methods

- Accuracy
  - classifier accuracy: predicting class label
  - predictor accuracy: guessing value of predicted attributes
- Speed
  - time to construct the model (training time)
  - time to use the model (classification/prediction time)
- Robustness: handling noise and missing values
- Scalability: efficiency in disk-resident databases
- Interpretability
  - understanding and insight provided by the model
- Other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules

Evaluation Criteria

- Accuracy on test set
  - The rate of correct classification on the testing set. E.g., if 90 are classified correctly out of the 100 testing cases, accuracy is 90%.
  - Actual evaluation in research work for several times.
- Error Rate on test set
  - The percentage of wrong predictions on test set.
- Confusion Matrix(混淆矩阵)
  - For binary class values, “yes” and “no” , a matrix showing true positive, true negative, false positive and false negative rates
- Speed and scalability
  - The time to build the classifier and to classify new cases, and the scalability with respect to the data size.
- Robustness: handling noise and missing values

Evaluation Criteria

		Predicted class	
		Yes	No
Actual class	Yes	True positive	False negative
	No	False positive	True negative

Evaluation Criteria

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: $\mathrm{Precision}(i) = \frac{R_{ii}}{\sum_{j=1}^{3} R_{ji}}$

7

Evaluation Techniques

- *Holdout*: the training set/testing set.
  - Good for a large set of data.
- *k-fold Cross-validation(交叉验证)*:
  - divide the data set into k sub-samples.
  - In each run, use one distinct sub-sample as testing set and the remaining k-1 sub-samples as training set.
  - Evaluate the method using the average of the k runs.
- This method reduces the randomness of training set/testing set.

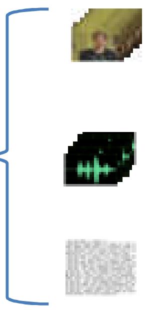
8

Issue 3: A Complete Classification Flow

- **Single Modal Information v.s. Multi-modal Information**
  - Single Modal Information
  - Multi-Modal Information

Issue 3: A Complete Classification Flow——
Unimodal Information and Multi-modal Information

- **Unimodal Information:** Data, text, audio (signal), video/picture, etc.
- **Typical Multimodal Information** :
  - **Short video:** TikTok/KuaiShou platform consumers' comments on products/services
  - **TCM "Four Diagnosis" information:** inspection (picture data), listen (audio information), question (text information), feel (pulse diagnosis-signal data)
  - **Other Multi-modal information:** gestures, postures, lip shapes, etc., here we focus on the information of different physical modalities



Video / Picture Information
-> Look(face, tongue, eye)

Audio Information
-> Listen (voice), pulse (signal) data

Text Information
-> Question (medical record) data

Issue 3: A Complete Classification Flow——Feature Engineering and Feature Learning Representation (1)

- **Classification** of unimodal information (typical problems of machine learning, function mapping problems)
  - Data Binarization Processing
    $f(x) = 1$ *or* -1
  - Speech Recognition
    $f($  $) =$ “Hello”
  - Image Processing
    $f($  $) =$ “9”
  - Smart Game(Go)
    $f($  $) =$ “6-5” (Placement position)
  - Machine Translation
    $f($ “你好！” $) =$ “Hello!”

11

Issue 3: A Complete Classification Flow——Feature Engineering and Feature Learning Representation (2)

- **Classification Mapping**
  - **Supervised learning classifier** (classification: traditional machine learning, deep neural network)
  - **Unsupervised learning classifier** (clustering)
  - **Semi-supervised learning classifier** (reinforcement learning problem: the case of small sample calibration data set)
- **How to obtain the characteristic description x of different modal information ?**
  - **Data Classification**
    - **Structured data:** Data, information in the database
    - **Semi-structured data:** News page content
    - **Unstructured data:** Pictures, videos (timing information), audio (timing information), etc.
  - **Feature Representation Method**
    - **Feature Engineering Method**
    - **Based on Learning Representation**：Learning into a feature space vector through a **data-driven mechanism**

12

Issue 3: A Complete Classification Flow——
Feature Engineering and Feature Learning Representation （3）

- **Features and Classification（1）**
  - Feature Engineering
    - **Text：**Letters, morphology, syntax, etc.
    - **Pictures：**Colors, textures, collection features, etc.
    - **Video**：Picture features + Temporal information
    - **Audio：**Signal features
  - Classification model based on feature engineering

特征工程（Feature Engineering）

13

Issue 3: A Complete Classification Flow——
Feature Engineering and Feature Learning Representation （4）

- **Features and Classification（2）**
  - **Learning Feature Representation and Classification**

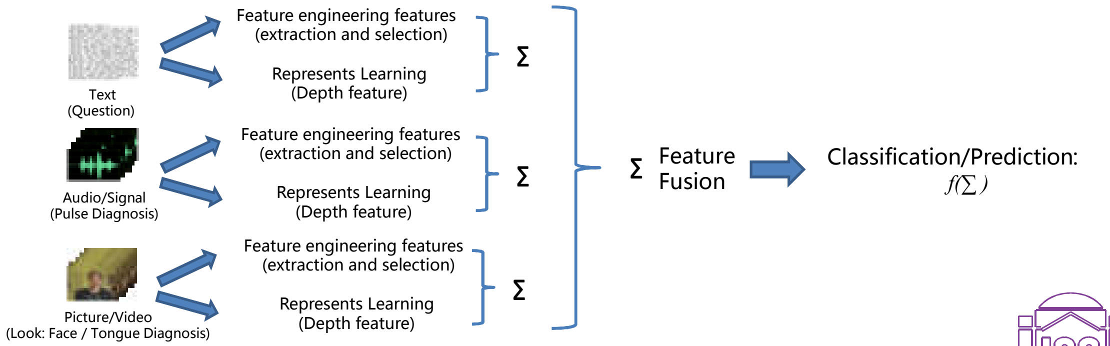
By building a model with a certain "depth", the model can automatically learn a good feature representation (from low-level features, to middle-level features, and then to high-level features), thereby ultimately improving the accuracy of prediction (classification) or recognition.

14

Issue 3: A Complete Classification Flow——
Feature Engineering and Feature Learning Representation （5）

Features and Classification（3）

- Which features of the unimodal information need to be fused? How to integrate? (Is the feature linear or vectorized?)
- At which level and which characteristics of multi-modal information need to be fused? How to integrate?
- How to determine the weight of the fused features according to the classification effect during the fusion process?



Thanks！