

THỰC HÀNH KHAI PHÁ DỮ LIỆU

Bài 2. Các mô hình khai phá dữ liệu trên weka

Giáo viên: TS. Trần Mạnh Tuấn

Bộ môn: Hệ thống thông tin

Khoa: Công nghệ thông tin

Email: tmtuan@tlu.edu.vn

Điện thoại: 0983.668.841

Nội dung

1

Giới thiệu về phân lớp dữ liệu

2

Giới thiệu về phân cụm dữ liệu

3

Giới thiệu về luật kết hợp

4

Giới thiệu về hồi quy dữ liệu

Giới thiệu về phân lớp dữ liệu

- ❖ **Mục đích:** để dự đoán nhãn phân lớp cho các bộ dữ liệu/mẫu mới
- ❖ **Đầu vào:** một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu
- ❖ **Đầu ra:** mô hình (bộ phân lớp) dựa trên tập huấn luyện và những nhãn phân lớp

Giới thiệu về phân lớp dữ liệu

Các bước phân lớp dữ liệu

- **Bước 1: Xây dựng mô hình từ tập huấn luyện:**
 - ✓ Mỗi bộ/mẫu dữ liệu được phân vào một lớp được xác định trước
 - ✓ Lớp của một bộ/mẫu dữ liệu được xác định bởi thuộc tính gán nhãn lớp
 - ✓ Tập các bộ/mẫu dữ liệu huấn luyện - tập huấn luyện tập huấn luyện được dùng để xây dựng mô hình
 - ✓ Mô hình được biểu diễn bởi các phương pháp phân lớp
- **Bước 2: Sử dụng mô hình - kiểm tra tính đúng đắn của mô hình và dùng nó để phân lớp dữ liệu mới:**
 - ✓ Phân lớp cho những đối tượng mới hoặc chưa được phân lớp
 - ✓ **Đánh giá độ chính xác của mô hình**
 - lớp biết trước của một mẫu/bộ dữ liệu đem kiểm tra được so sánh với kết quả thu được từ mô hình
 - tỉ lệ chính xác = phần trăm các mẫu/bộ dữ liệu được phân lớp đúng bởi mô hình trong số các lần kiểm tra

Các mô hình phân lớp dữ liệu

- **Cây quyết định**
 - **Naïve Bayes**
 - **Mô hình thống kê**
 - **Mạng nơ ron**
 - **Mô hình SVM**
 - **Mô hình KNN**
 - **Các mô hình khác**
-

Phân lớp dữ liệu trên weka

- ❖ Là một chức năng của Explorer
- ❖ Hỗ trợ người dùng huấn luyện và kiểm chứng các mô hình phân lớp cơ bản

Các bước thực hiện phân lớp dữ liệu

- ❖ **Bước 1: tại tab Preprocess, chọn tập dữ liệu và tiền xử lý dữ liệu**
- ❖ **Bước 2: Chọn thuật toán phân lớp và xác định tham số**
- ❖ **Bước 3: Chọn kiểu test và tập dữ liệu test (nếu cần)**
- ❖ **Bước 4: Tiến hành phân lớp dữ liệu**
- ❖ **Bước 5: Ghi nhận và phân tích kết quả**

Giới thiệu về phân lớp dữ liệu

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **J48 -C 0.25 -M 2**

Test options

☐ Use training set

☐ Supplied test set **Set...**

☒ Cross-validation Folds **10**

☐ Percentage split % **66**

More options...

(Nom) class

Start **Stop**

Result list (right-click for options)

10:14:04 - trees.J48

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      144           96   %
Incorrectly Classified Instances     6            4   %
Kappa statistic                    0.94
Mean absolute error                 0.035
Root mean squared error             0.1586
Relative absolute error             7.8705 %
Root relative squared error        33.6353 %
Total Number of Instances         150

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.980	0.000	1.000	0.980	0.990	0.985	0.990	0.987	Iris-setosa
	0.940	0.030	0.940	0.940	0.940	0.910	0.952	0.880	Iris-versicolor
	0.960	0.030	0.941	0.960	0.950	0.925	0.961	0.905	Iris-virginica
Weighted Avg.	0.960	0.020	0.960	0.960	0.960	0.940	0.968	0.924	

```
=== Confusion Matrix ===

 a  b  c  <-- classified as
49  1  0 | a = Iris-setosa
 0 47  3 | b = Iris-versicolor
 0  2 48 | c = Iris-virginica
```

Status

Log

Chọn kiểu test phân lớp dữ liệu

- ❖ Sử dụng chính tập huấn luyện làm tập test: use training set
- ❖ Chỉ định tập test mới: supplied test set
- ❖ Chia tỉ lệ test theo k-folds: Cross validation
- ❖ Chia tỷ lệ phần trăm trên data: Percentage split
- ❖ Các lựa chọn chỉnh sửa khác: more options

Giới thiệu về phân lớp dữ liệu

Kết quả phân lớp dữ liệu

Classifier output

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: iris

Instances: 150

Attributes: 5

sepallength

sepalwidth

petallength

petalwidth

class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

Giới thiệu về phân lớp dữ liệu

Kết quả phân lớp dữ liệu

- ❖ **Classifier mode (full training set):** cho biết mô hình phân lớp dựa trên cả tập huấn luyện, cây quyết định, thời gian chạy mô hình

```
--- Classifier model (full training set) ---  
  
J48 pruned tree  
-----  
  
petalwidth <= 0.6: Iris-setosa (50.0)  
petalwidth > 0.6  
|   petalwidth <= 1.7  
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)  
|   |   petallength > 4.9  
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)  
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)  
|   |   petalwidth > 1.7: Iris-virginica (46.0/1.0)  
  
Number of Leaves      :          5  
  
Size of the tree      :          9  
  
Time taken to build model: 0.01 seconds
```

Giới thiệu về phân lớp dữ liệu

Kết quả phân lớp dữ liệu

- ❖ Tổng kết: số liệu thống kê cho biết độ chính xác của bộ phân lớp, theo kiểu test cụ thể.

```
=== Stratified cross-validation ===  
=== Summary ===
```

Correctly Classified Instances	144	96	%
Incorrectly Classified Instances	6	4	%
Kappa statistic	0.94		
Mean absolute error	0.035		
Root mean squared error	0.1586		
Relative absolute error	7.8705	%	
Root relative squared error	33.6353	%	
Total Number of Instances	150		

Kiểu test

Số mẫu
phân
lớp
đúng

Số mẫu
phân
lớp
sai

Các thông số
khác

Giới thiệu về phân lớp dữ liệu

Kết quả phân lớp dữ liệu

- ❖ **Độ chính xác của từng phân lớp với các độ đo phân lớp:**

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.980	0.000	1.000	0.980	0.990	0.985	0.990	0.987	Iris-setosa
	0.940	0.030	0.940	0.940	0.940	0.910	0.952	0.880	Iris-versicolor
	0.960	0.030	0.941	0.960	0.950	0.925	0.961	0.905	Iris-virginica
Weighted Avg.	0.960	0.020	0.960	0.960	0.960	0.940	0.968	0.924	

Giới thiệu về phân lớp dữ liệu

Kết quả phân lớp dữ liệu

- ❖ **Confusion Matrix:** cho biết bao nhiêu mẫu được gán vào từng lớp. Các phần tử của ma trận thể hiện số mẫu test có lớp thật sự là dòng, lớp dự đoán là cột

```
=== Confusion Matrix ===
```

```
  a   b   c   <-- classified as
49   1   0   |   a = Iris-setosa
 0  47   3   |   b = Iris-versicolor
 0   2  48   |   c = Iris-virginica
```

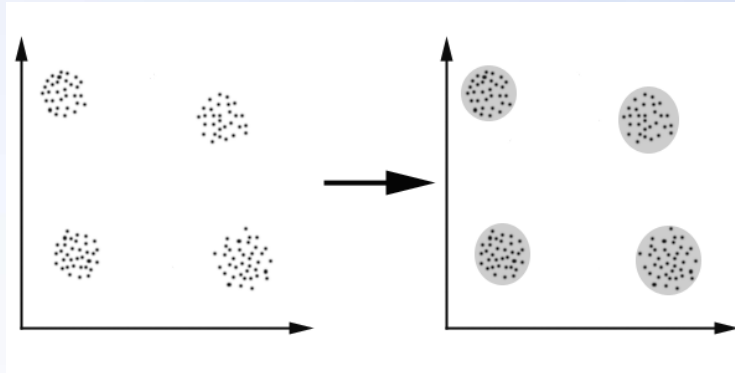
Giới thiệu về phân lớp dữ liệu

Tổng hợp so sánh phân lớp dữ liệu

- ❖ Chạy trên cùng 1 bộ dữ liệu: Iris
- ❖ Phương pháp:
 - Cây quyết định J48, RadoForest
 - Naïve Bayes
 - AdaBoostM1
 - LWL
 - Jrip

Giới thiệu về phân cụm dữ liệu

Phân cụm dữ liệu



- ❖ **Phân cụm rõ:** các điểm dữ liệu được chia vào các cụm, trong đó mỗi điểm dữ liệu thuộc vào chính xác một cụm.
- ❖ **Phân cụm mờ:** các điểm dữ liệu có thể thuộc vào nhiều hơn một cụm với độ thuộc tương ứng.

Giới thiệu về phân cụm dữ liệu

Phân cụm dữ liệu trên weka

- ❖ Là một chức năng của Explorer
- ❖ Hỗ trợ người dùng huấn luyện và kiểm chứng các mô hình phân cụm cơ bản

Giới thiệu về phân cụm dữ liệu

Các bước thực hiện phân lớp dữ liệu

- ❖ **Bước 1: tại tab Preprocess, chọn tập dữ liệu và tiền xử lý dữ liệu**
- ❖ **Bước 2: Chọn thuật toán phân cụm và xác định tham số**
- ❖ **Bước 3: Chọn tập phân cụm**
- ❖ **Bước 4: Tiến hành phân cụm dữ liệu**
- ❖ **Bước 5: Ghi nhận và phân tích kết quả**

Giới thiệu về phân cụm dữ liệu

Weka Explorer

Preprocess | **Cluster** | Associate | Select attributes | Visualize

Clusterer

Choose **SimpleKMeans** -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

- ☒ Use training set
- ☐ Supplied test set
- ☐ Percentage split %
- ☐ Classes to clusters evaluation (Num) c
- ☒ Store clusters for visualization

Result list (right-click for options)

22:26:59 - SimpleKMeans

Clusterer output

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Cluster#	
	0	1
	(768.0)	(515.0)
preg	3.8451	2.0835
plas	120.8945	115.3282
pres	69.1055	65.9903
skin	20.5365	21.8194
insu	79.7995	85.0194
mass	31.9926	31.7751
pedi	0.4719	0.4708
age	33.2409	26.7728


Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

Cluster	Count	Percentage
0	515	67%
1	253	33%

Status

 x 0

Giới thiệu về phân cụm dữ liệu

Tổng hợp so sánh phân cụm dữ liệu

- ❖ Chạy 1 bộ dữ liệu với các phương pháp phân cụm khác nhau
- ❖ Chạy thuật toán K-mean với các bộ dữ liệu khác nhau

➤ Giới thiệu về luật kết hợp

Khai phá luật kết hợp:

- Tìm tần số mẫu, mối kết hợp, sự tương quan, hay các cấu trúc nhân quả giữa các tập đối tượng trong các cơ sở dữ liệu giao tác, cơ sở dữ liệu quan hệ, và những kho thông tin khác.

Tính hiểu được: dễ hiểu

Tính sử dụng được: Cung cấp thông tin thiết thực

Tính hiệu quả: Đã có những thuật toán khai thác hiệu quả

Các ứng dụng:

- Phân tích bán hàng trong siêu thị, cross-marketing, thiết kế catalog, loss-leader analysis, gom cụm, phân lớp, ...
-

➤ Giới thiệu về luật kết hợp

Các khái niệm

Cho $I = \{I_1, I_2, \dots, I_m\}$ là tập các đơn vị dữ liệu. Cho D là tập các giao tác, mỗi giao tác T là tập các đơn vị dữ liệu sao cho $T \subseteq I$

Định nghĩa 1: Ta gọi giao tác T chứa X , với X là tập các đơn vị dữ liệu của I , nếu $X \subseteq T$

Định nghĩa 2: Một luật kết hợp là một phép suy diễn có dạng $X \rightarrow Y$, trong đó $X \subset I$, $Y \subset I$ và $X \cap Y = \emptyset$

Định nghĩa 3: Ta gọi luật $X \rightarrow Y$ có mức xác nhận(support) là s trong tập giao tác D , nếu có $s\%$ giao tác trong D chứa $X \cup Y$.
Ký hiệu: $\text{Supp}(X \rightarrow Y) = s$

➤ Giới thiệu về luật kết hợp

Định nghĩa 4: Ta gọi luật $X \rightarrow Y$ là có độ tin cậy c (Confidence) trên tập giao tác D ,

Ký hiệu: $c = \text{Conf}(X \rightarrow Y) = \text{Supp}(X \rightarrow Y) / \text{Supp}(X)$

Nhận xét: Các xác nhận và độ tin cậy chính là các xác suất sau:

$\text{Supp}(X \rightarrow Y) = P(X \cup Y)$: Xác suất của $X \cup Y$ trong D

$\text{Conf}(X \rightarrow Y) = P(Y/X)$: Xác suất có điều kiện

Định nghĩa 5: Cho trước $\text{Min_Supp} = s_0$ và $\text{Min_Conf} = c_0$

Ta gọi luật $X \rightarrow Y$ là xảy ra nếu thỏa:

$\text{Supp}(X \rightarrow Y) > s_0$ và $\text{Conf}(X \rightarrow Y) > c_0$

➤ Giới thiệu về luật kết hợp

- Thuật toán Apriori
- Thuật toán FP-growth

➤ Giới thiệu về luật kết hợp

Luật kết hợp trên weka

- ❖ Là một chức năng của Explorer
- ❖ Hỗ trợ người dùng huấn luyện và kiểm chứng các thuật toán luật kết hợp cơ bản

➤ Giới thiệu về luật kết hợp

Các bước thực hiện luật kết hợp

- ❖ **Bước 1:** tại tab Preprocess, chọn tập dữ liệu và tiền xử lý dữ liệu: các trường dữ liệu dạng Nominal. Nếu ở dạng khác thì dùng bộ lọc để chuyển về: NumericToNominal
- ❖ **Bước 2:** Chọn thuật toán luật kết hợp và tham số
- ❖ **Bước 3:** Tiến hành thực hiện thuật toán
- ❖ **Bước 4:** Ghi nhận và phân tích kết quả

Giới thiệu về luật kết hợp

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Associator

Choose **Apriori -N 50 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.4 -S -1.0 -c -1**

Start Stop

Result list (right-click...)

21:28:00 - Apriori

Associator output

```
=== Run information ===

Scheme:      weka.associations.Apriori -N 50 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.4 -S -1.0 -c -1
Relation:    supermarket-weka.filters.unsupervised.attribute.Remove-R1-9,11,15-weka.filters.unsupervised.attribute.F
Instances:   4627
Attributes:  107
              [list of attributes omitted]

=== Associator model (full training set) ===

Apriori
=====

Minimum support: 0.4 (1851 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 17

Size of set of large itemsets L(2): 16

Best rules found:

1. biscuits=t 2605 ==> bread and cake=t 2083    <conf:(0.8)> lift:(1.11) lev:(0.04) [208] conv:(1.4)
2. milk-cream=t 2939 ==> bread and cake=t 2337    <conf:(0.8)> lift:(1.1) lev:(0.05) [221] conv:(1.37)
3. fruit=t 2962 ==> bread and cake=t 2325    <conf:(0.78)> lift:(1.09) lev:(0.04) [193] conv:(1.3)
4. baking needs=t 2795 ==> bread and cake=t 2191    <conf:(0.78)> lift:(1.09) lev:(0.04) [179] conv:(1.29)
```

➤ Giới thiệu về luật kết hợp

Tổng hợp so sánh luật kết hợp

- ❖ Chạy 1 bộ dữ liệu với các phương pháp thuật toán khác nhau
- ❖ Chạy thuật toán Apriori với các bộ dữ liệu khác nhau

Giới thiệu về Hồi quy dữ liệu

- Chủ yếu dùng để dự đoán đầu ra (định lượng)
- Đầu vào và đầu ra có mối quan hệ dưới dạng 1 hàm bậc nhất (tuyến tính):

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Trong đó:

α là hệ số chặn; β là độ dốc (hệ số hồi quy)

ε_i là một biến số theo luật phân phối chuẩn

Giới thiệu về Hồi quy dữ liệu

- Mô hình chỉ có 1 biến dùng để dự đoán biến đích
 - Dễ dàng xác định được đường thẳng “phù hợp nhất”
-

Giới thiệu về Hồi quy dữ liệu

➤ Trong mô hình:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Các hệ số α và β được xác định theo phương pháp bình phương cực tiểu

Trao đổi, câu hỏi?