Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Matthias Schubert, Dr. Eirini Ntoutsi
Erich Schubert

Knowledge Discovery in Databases II

WS 2012/2013

Übungsblatt 8: Ensemble Learning und Multi-Repräsentierte Daten

Aufgabe 8-1 *Error Correcting Output Codes*

(a) Beschreiben Sie das Klassifikationsschema *one-versus-rest* für ein 4-Klassen-Problem in der Notation, die Sie für ECOCs kennengelernt haben.

(b) Beschreiben Sie ein ECOC-Schema für eine minimale Anzahl von Base-Classifiern an sowie ein vollständiges ECOC-Schema, das die Codes für jede nicht-trivialen Aufteilung der 4 Klassen in ein zwei-elementige Menge von Klassen angibt.

Was beobachten Sie für die Row-Separation?

Aufgabe 8-2 *Ensemble Multi-Klassen-Klassifikation*

In der Vorlesung haben Sie die Ensemble-Techniken *one-versus-rest*, *all-pairs* und *ECOC* kennengelernt, die Klassifikationsprobleme mit mehr als 2 Klassen auf mehrere 2-Klassen-Probleme zurückführen. Für *one-versus-rest* und *all-pairs* können wir in der Test-/Anwendungsphase ein einfaches Mehrheitsvoting für die Klassifikationsentscheidung annehmen. Für *ECOC* haben wir die Entscheidungsregel genauer diskutiert. Eine weitere Möglichkeit stellt das DDAG-Schema dar: Aus den einzelnen *all-pairs*-Klassifikatoren wird ein gerichteter, azyklischer Graph für die Klassifikationsentscheidung gebildet (DDAG=Decision Directed Acyclic Graph), siehe Abbildung 1.

(a) Welche Vor- oder Nachteile hat diese Strategie gegenüber dem Voting über alle paarweisen Klassifikatoren?

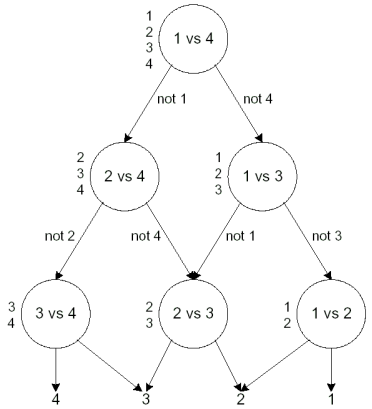

Abbildung 1: Klassifikationsschema DDAG

(b) Nehmen Sie als Komplexität eines Base-Classifiers im Training die Funktion $t : \mathbb{N} \to \mathbb{R}_0^+$ an, die abhängig von der Anzahl der Trainingsbeispiele ist. Wie verhalten sich die unterschiedlichen Schemata hinsichtlich ihres Zeitbedarfs in der Trainingsphase bei n Klassen und m Beispielen für jede Klasse? Wie sieht es in der Anwendungsphase aus, wenn Sie einen konstanten Zeitbedarf für die Vorhersage des einzelnen Base-Klassifikators annehmen?

Aufgabe 8-3 *Kombination von zwei Ähnlichkeitsmaßen*

Gegeben seien zwei Kernels k_1 und k_2. Wir kombinieren sie in einen gemeinsamen Kernel k_{com}

$$k_{com} = \alpha k_1 + (1 - \alpha)k_2, \tag{1}$$

wobei $\alpha \in [0; 1]$.

Wir wenden den Kernel k_{com} auf zwei Klassifikationsprobleme an, wobei wir das Experiment für verschiedene Werte von α wiederholen. In der folgenden Abbildung stellen wir die Klassifikationsgenauigkeit auf dem ersten Datensatz (R1) und auf dem zweiten Datensatz (R2) in Abhängigkeit von α dar:

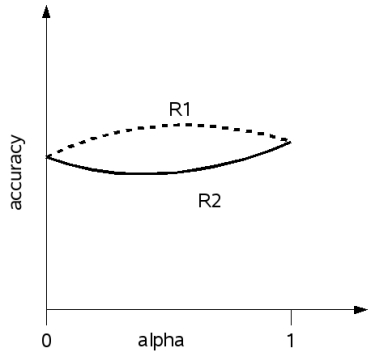

Abbildung 2: Classification accuracy vs. α

Beantworten Sie folgende Fragen anhand von Abbildung 2:

(a) Auf welchem der beiden Datensätze lohnt sich die Kombination der beiden Kernels?

(b) Wann funktionieren die Kernels k_1 und k_2 alleine besser als kombiniert?

Aufgabe 8-4 *Komplementarität von Klassifikatoren*

Gegeben seien zwei binäre Klassifikatoren f_1 und f_2, die auf je einer Repräsentation der Objekte eines Datensatzes D mit Klassen $\{0, 1\}$ arbeiten. Entscheiden Sie, ob in den folgenden Fällen eine Kombination der Klassifikatoren sinnvoll ist:

(a) $f_1(x) = f_2(x)$ für alle $x \in D$

(b) $f_1(x) = 1 - f_2(x)$ für alle $x \in D$

Aufgabe 8-5 *Abhängigkeitsmaß*

Gegeben sei ein Maß h, welches die Abhängigkeit zwischen zwei Kernelmatrizen K und K' misst. Anschaulich heißt das, dass $h(K, K')$ groß ist, wenn die zugehörigen Kernels k und k' dieselben Objekte als ähnlich und als unähnlich betrachten. Wenn sie die Ähnlichkeit derselben Objekte unterschiedlich bewerten, sei $h(K, K')$ niedrig.

Seien nun ein Datensatz D mit einem Klassenlabel und r Repräsentationen pro Objekt gegeben. Wir berechnen eine Kernelmatrix K_i für jede der r Repräsentationen und eine Kernelmatrix L auf den Klassenlabels. Überlegen Sie sich, wie man mittels h eine Linearkombination der K_i bestimmen kann, die die Ähnlichkeit der Klassenlabels möglicht gut widerspiegelt.

Aufgabe 8-6 *Multirepräsentiertes Clustering*

Gegeben sei ein Datensatz X, so dass jeder Punkt durch 2 zweidimensionale Vektoren repräsentiert wird.

```
A = ( 0, 1); ( 3, 0)     B = (-1,-1); ( 2, 0)     C = ( 0, 0); ( 3, 1)
D = ( 0,-3); (-2, 2)     E = ( 2, 1); (-2,-3)
```

Wir wollen auf diesem Datensatz multirepräsentiertes Clustering mittels DBSCAN durchführen.

(a) Wie unterscheidet sich multirepräsentiertes Clustering von gewöhnlichem Clustering? Welche besonderen Schwierigkeiten sind damit verbunden?

(b) Es sei $MinPoints = 3$. Für welche Werte von $\varepsilon_1, \varepsilon_2$ sind die Objekte C und D Kernobjekte nach

- der Vereinigungsmethode?
- der Schnittmethode?

Aufgabe 8-7 *Multi-Repräsentierte Klassifikation*

Gegeben sei ein Datensatz mit multiplen Repräsentationen jedes Datenobjekts. Wir möchten Klassenzugehörigkeiten mittels dieser multiplen Repräsentationen ermitteln.

- In welcher Phase des Klassifikationsprozesses können wir die verschiedenen Repräsentationen integrieren?
- Wie können wir die multiplen Repräsentationen beim Trainieren integrieren?
- Wie können wir die multiplen Repräsentationen beim Vorhersagen integrieren?
- Ist in beiden Fällen zuvor eine Normalisierung erforderlich?