9. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 12.01.2015, 18:00 Uhr

WS 2014/2015
Prof. W. Rhode

Aufgabe 30: *Data-Mining Challenge* **20 P.**

In dieser Aufgabe soll eine reale Problemstellung, mit der sich unser Lehrstuhl aktuell beschäftigt, behandelt werden. Es handelt sich dabei um eine Aufgabe innerhalb des FACT-Experiments.

Problemstellung: Das FACT-Experiment hat das Ziel Gammastrahlung aus dem Universum von der Erde aus zu messen. Diese Gammastrahlung kann jedoch nicht direkt gemessen werden, da die Gammastrahlung nicht die Erdatmosphäre durchdringen kann. Allerdings löst ein Gamma einen Teilchenschauer in der Atmosphäre aus, in dem Cherenkov-Licht produziert wird, welches mit speziellen Teleskopen gemessen werden kann. In der Kamera des Teleskops erscheint der Cherenkov-Lichtkegel als Ellipse, die einen Rückschluss auf die Energie und die Richtung des Gammas erlaubt. Jedoch löst nicht nur Gammastrahlung Teilchenschauer aus, sondern auch geladene Teilchen der primären kosmischen Strahlung. (Diese besteht größtenteils aus Protonen, weshalb im folgenden von Protonen geredet wird.) Aufgrund der unterschiedlichen Teilchen und Wechselwirkungen sehen die Teilchenschauer und die daraus resultierenden Ellipsen unterschiedlich aus. Die Trennung der Gamma-Ereignisse von den Proton-Ereignissen ist ein aktuelles Problem in der Forschung.

Ein Data-Mining Ansatz ist prädestiniert zu dessen Lösung und wir bitten Euch um Eure Mithilfe! ;-)

Datensatz: Die Daten sind bereits vorverarbeitet, so dass Sie eine gewisse Menge an Attributen, die die Ellipse und das gesamte Kamerabild beschreiben, gegeben haben. Eine genauere Beschreibung der einzelnen Attribute finden Sie am Ende der Aufgabe. Im EWS finden Sie folgende Dateien:

Gamma-MC.csv In dieser Datei finden Sie MCs von Gammas aus einer Quellrichtung, auf die das Teleskop gerichtet ist.

Proton-MC.csv In dieser Datei finden Sie MCs von Protonen aus allen Richtungen.

Test-Daten.csv In dieser Datei finden Sie sowohl MCs von Gammas als auch von Protonen in einem Verhältnis von $1:10$ (Gamma:Proton), die als Test-Daten verwendet werden sollen.

Es steht Ihnen frei, wie viel Zeit und Anstrengung Sie für diese Aufgabe verwenden, es wird jedoch **Buchpreise und Urkunden für die beste Trennung** geben.

Die Qualität der Trennung werden wir mithilfe des sogenannten Q-Faktors bestimmen, der wie folgt definiert ist:

$$Q = \frac{E_{Gamma}}{\sqrt{E_{Proton}}} \tag{1}$$

Die Effizienzen für Gamma und Protonen berechnen sich wie folgt:

$$E = \frac{N_{tp}}{N_{tp} + N_{fn}}$$

a) Erstellen Sie einen Prozess im RapidMiner, der ein Modell zur Trennung von Gammas und Protonen erzeugt. Die Wahl des Lerners steht Ihnen frei. **5 P.**

b) Wenden Sie dieses Modell auf den Test-Datensatz an. Wählen Sie einen geeigneten Konfidenz-Schnitt (optimierbar nach Gleichung 1), wenden Sie diesen an und klassifizieren Sie den Datensatz, indem Sie jedem Ereignis ein Label (1 für Gamma, 0 für Proton) hinzufügen. **5 P.**

c) Berechnen Sie den Q-Faktor zur Bewertung Ihrer Performanz in Abhängigkeit verschiedener Konfidenz-Schnitte und stellen Sie dies in einem Plot dar (einfacher außerhalb von RapidMiner mit Programm Ihrer Wahl). **2 P.**

d) Validieren Sie Ihren Prozess mit einer Kreuzvalidierung und fügen Sie den so erhaltenen Fehler auf den Q-Faktor dem Plot hinzu. **2 P.**

e) Zur Verbesserung Ihres Prozesses führen Sie eine Feature Generation durch. **2 P.**
(Ein anschließender Vergleich der Q-Faktoren kann Ihnen Hinweise zur Optimierung geben. Passen Sie Ihren obigen Prozess und Ihren Konfidenz-Schnitt entsprechend an.)

f) Zur Verbesserung Ihres Prozesses führen Sie eine Feature Selection mithilfe des MRMR-Algorithmus' durch. **2 P.**
(Ein anschließender Vergleich der Q-Faktoren kann Ihnen Hinweise zur Optimierung geben. Passen Sie Ihren obigen Prozess und Ihren Konfidenz-Schnitt entsprechend an.)

g) Optimieren Sie die Einstellungen Ihres Lerners. **1 P.**
(Ein anschließender Vergleich der Q-Faktoren kann Ihnen Hinweise zur Optimierung geben. Passen Sie Ihren obigen Prozess und Ihren Konfidenz-Schnitt entsprechend an.)

h) Variieren Sie Ihr Trainigsverhältnis. Achten Sie darauf, dass Sie genügend Ereignisse für eine Klasse haben. **1 P.**
(Ein anschließender Vergleich der Q-Faktoren kann Ihnen Hinweise zur Optimierung geben. Passen Sie Ihren obigen Prozess und Ihren Konfidenz-Schnitt entsprechend an.)

9. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 12.01.2015, 18:00 Uhr

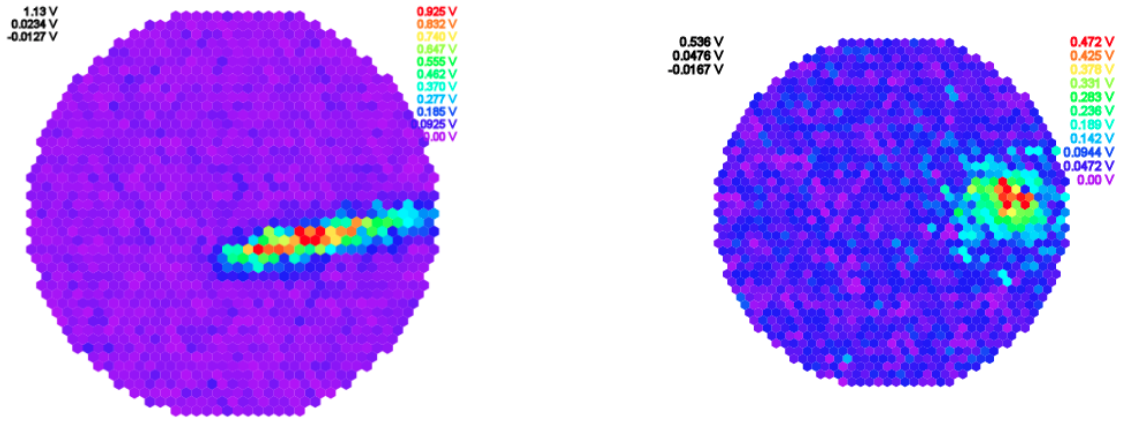
WS 2014/2015
Prof. W. Rhode

Abgabe: Schicken Sie uns bitte Folgendes:

- Den RapidMiner-Prozess.
- Die Datei Test-Daten.csv inklusive Ihrer ermittelten Label (und IDs).
- Ihre Plots der verschiedenen Q-Faktoren (bitte mit den gleichen x- und y-Bereichen).

Hinweise:

- Schauen Sie sich den Operator `Set Role` an.
- Achten Sie auf die Werte und den Type ihres Labels.
- Sie können bereits beim Einlesen Ihre ID als ID kennzeichnen.
- Die Konfidenz für die einzelnen Klassen werden vom jeweiligen Lerner zurückgegeben.



Kamerabild eines Gamma-Ereignisses (links) und eines Proton-Ereignisses (rechts).

Beschreibung der Attribute:

photonchargeMean Mittelwert über alle Photonen (Ladungen) im Kamerabild (Schauer-Photonen und Untergrund-Photonen von Umgebungslicht etc.)

arrivalTimeMean Mittelwert über die Ankunftszeiten der Cherenkov-Photonen der Ellipse

phChargeShower_mean Mittelwert über alle Photonen innerhalb der Ellipse

phChargeShower_kurtosis Kurtosis (= 4. Moment) aller Photonen innerhalb der Ellipse

phChargeShower_variance Varianz (= 2. Moment) aller Photonen innerhalb der Ellipse

phChargeShower_skewness Schiefe (= 3. Moment) aller Photonen innerhalb der Ellipse

arrTimeShower_skewness Schiefe (= 3. Moment) aller Photon-Ankunftszeiten innerhalb der Ellipse

numPixelInShower Anzahl der Pixel innerhalb der Ellipse

Size Summe aller Photonen innerhalb der Ellipse

Length 2. Moment aller Photonen entlang der langen Ellipsenachse

Width 2. Moment aller Photonen entlang der kurzen Ellipsenachse

numIslands Anzahl der entstandenen Pixelgruppen, die Photonen enthalten

Delta Drehwinkel des Schauers

COGx Schwerpunkt in x-Richtung der Kamera

COGy Schwerpunkt in y-Richtung der Kamera

M3Long 3. Moment entlang der langen Schauerachse

M3Trans 3. Moment entlang der kurzen Schauerachse

M4Long 4. Moment entlang der langen Schauerachse

M4Trans 4. Moment entlang der kurzen Schauerachse

Disp Distanz zwischen Schauer-Schwerpunkt und rekonstruierter Quellposition in der Kamera

Concentration_onePixel Quotient vom Pixel mit der höchsten Ladung und der Gesamtladung der Ellipse

Concentration_twoPixel Quotient der zwei Pixel mit den höchsten Ladungen und der Gesamtladung der Ellipse

ConcCore Quotient der Pixel innerhalb der theoretischen Ellipse und der Gesamtladung der Ellipse

concCOG Quotient der drei Pixel in der Nähe vom Schwerpunkt und der Gesamtladung des Schauers

Slope_long Steigung des Fits im x-t-Raum

Slope_trans Steigung des Fits im y-t-Raum

Timespread Breite der Zeitverteilung

Timespread_weighted Breite der Zeitverteilung gewichtet mit der Photon-Ladung

Alpha Winkel zwischen Hauptachse der Ellipse und der Strecke zwischen Schwerpunkt und theoretischer Quellposition in der Kamera

Distance Distanz zwischen Schwerpunkt und rekonstruierter Quellposition in der Kamera

CosDeltaAlpha Cosinus von Delta mal Alpha

Theta Distanz zwischen theoretischer Quellposition und rekonstruierter Quellposition in der Kamera