Blatt 9

Aufgabe 1

Die Daten werden mit dem Prozess namens 'retrieval.rmp' eingelesen und als Datensatz im lokalen repsoitory gespeichert. Folgende Prozesse bezuehen die Daten direkt aus dem Repository. Aus effizienz Gründen wurde häufig nur ein Teilmenge der Daten betrachtet. (Siehe den Stratified Sampling Operator)

Hinweise:

- Die Namen der Label lauten nicht wie in der Aufgabenstellung 0 und 1 sondern "proton" und "gamma".
- Es wurde RapidMiner Version 6.20 verwendet.
- Die Formel für den Q-Faktor wurde durch die Korrekte ersetzt. Da such der korrigierte Zettel offenbar vermurkst war.
- Aufgrund von zu kleinem Prozessor wurden nur sehr kleine Stichproben verwendet und nur 5-Fache Kreuzvalidierung. Dadurch werden die Fehler an den Q-Plots entsprechend Größer.

a)

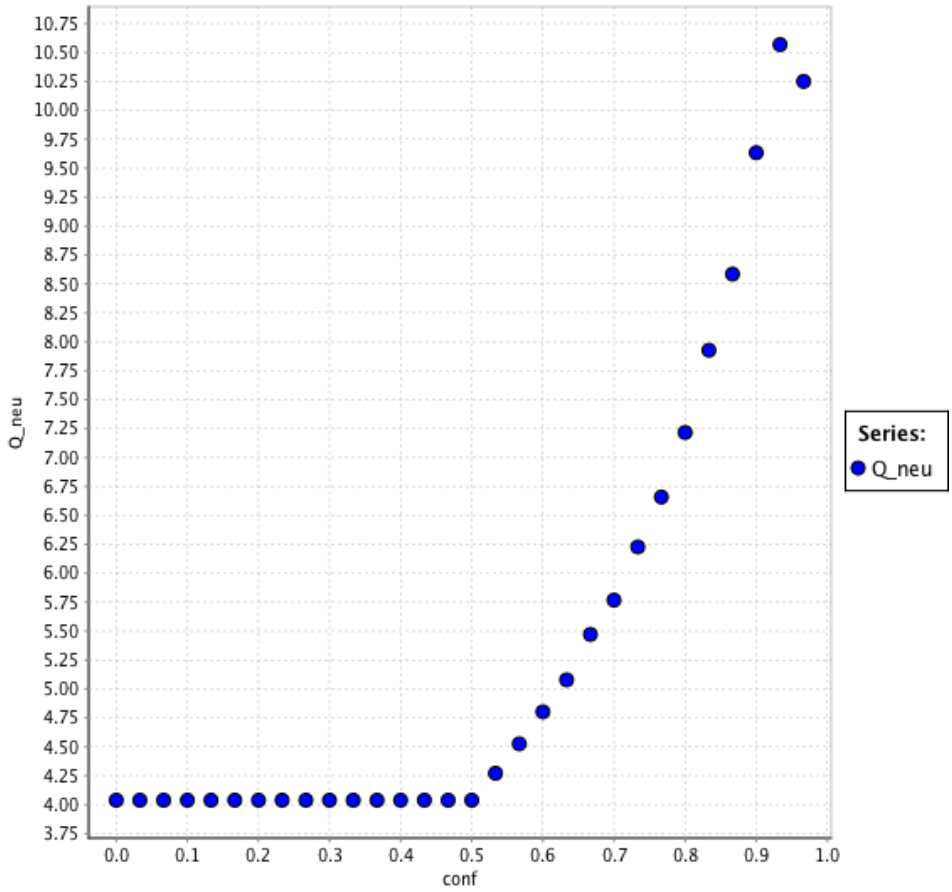
Datei 'aufgabe_a.rmp'. Als Klassifizierer wurde der Weka-Random Forest gewählt.

b)

Das Modell wird zunächst trainiert und dann mit dem Apply Modell Operator auf die ungelabelten Testdatenangewandt. Ein Konfidenzschnitt macht bei Anwendung des Modells auf ungelabelte Daten zunächst keinen Sinn. Dies geschieht in Aufgabe c)

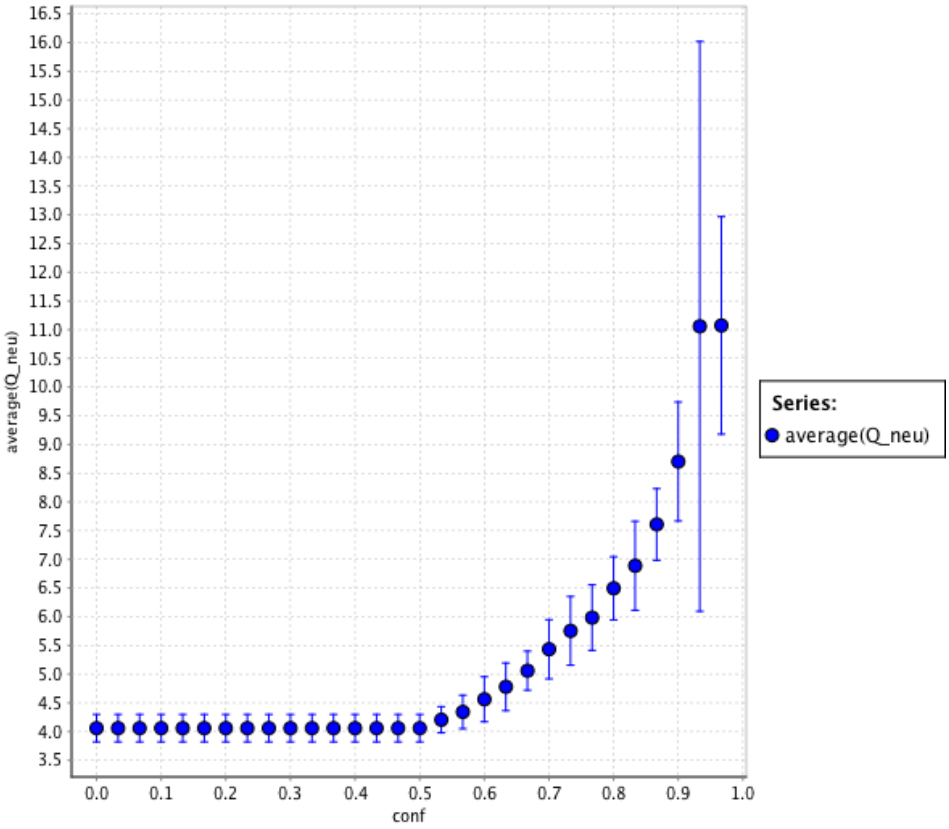
c)

Der confidence cut wird mittels des 'Drop Uncertain Prediction' Operators und anschließendem Auffüllen fehlender Werte durchgeführt. Der Wert für den Confidence Cut wird in einem Loop Operator angepasst. Die Werte aus dem Performance Vector werden mit einem Log Operator gespeichert und später in ein exampleset umgewandelt um die Q-Werte zu



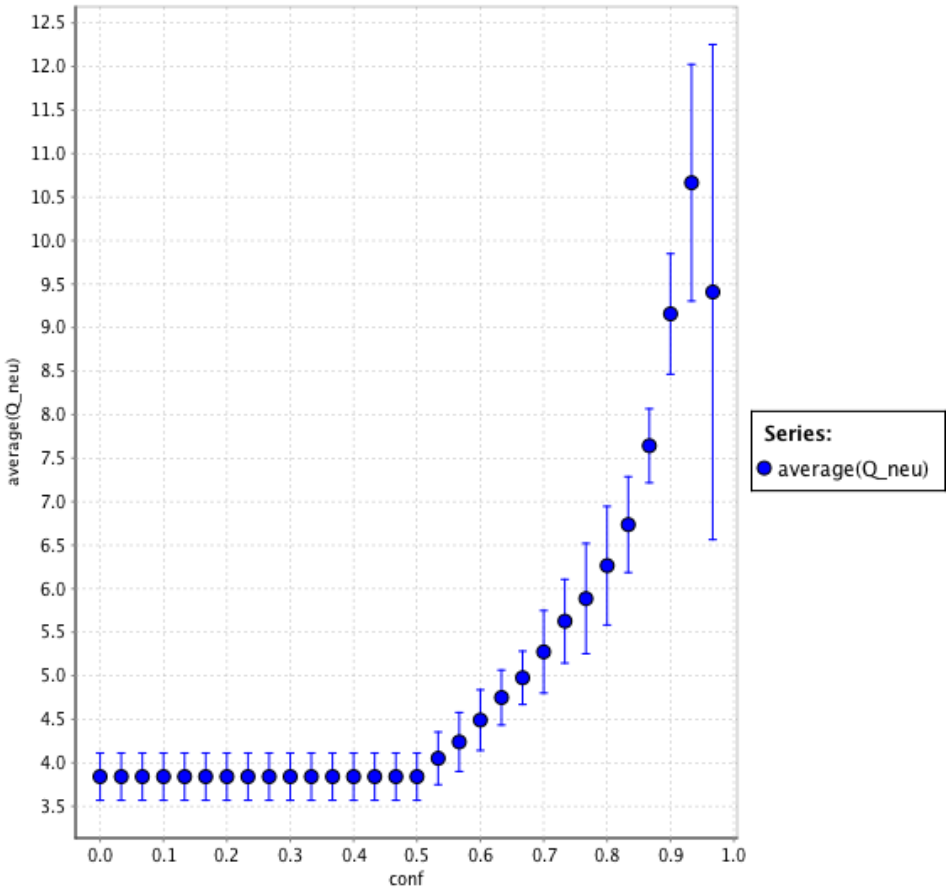
d)

Der Prozess ist Ähnlich zu dem aus Aufgabe c). Nur wird diesmal der Classifier und der Loop über die Konfidenzen innerhalb der Kreuzvalidierung hinzugefügt.



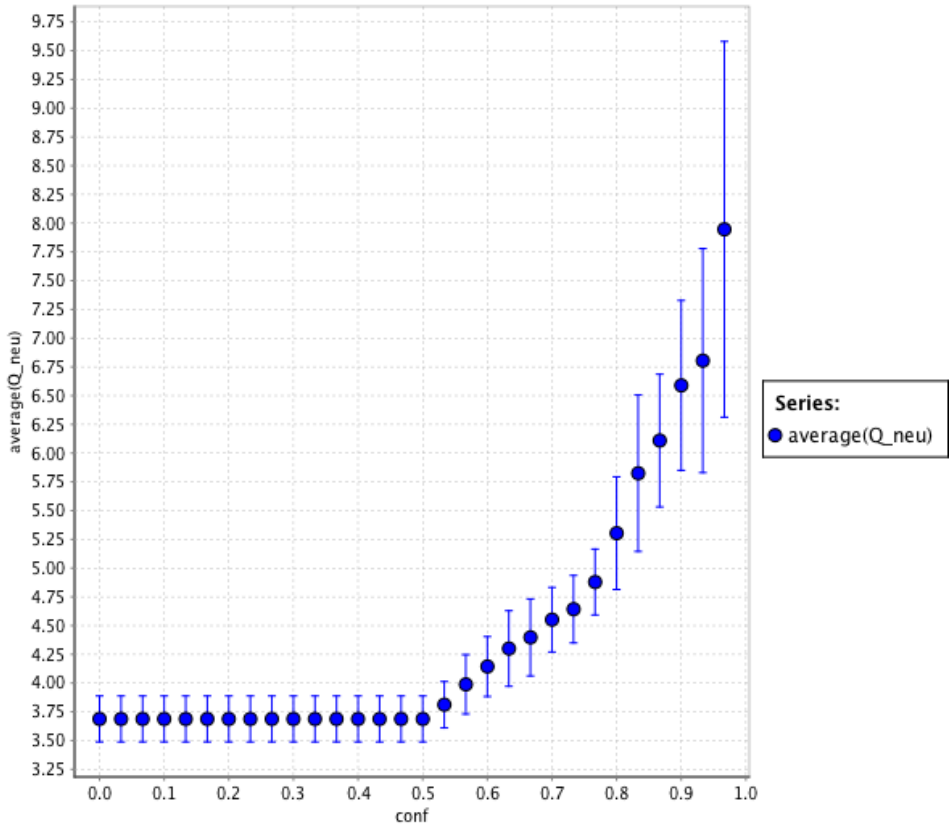
e)

Dasselbe in grün nur mit Feature Selection. Aus Speicherplatzgründen wurden nur zwei neue Kombinationen ausgewählt.



f)

Eine Selection mit MRMR macht bei Nutzung eines RandomForest erst dann Sinn wenn man wesentlich mehr Features als Trees benutzt. Wie erwartet wird die Trennung schlechter.



g) und h)

Zur Optimierung kann die Anzahl der Bäume des Lerners erhöht werden. Dadurch wird sowohl die Klassifizierung als auch die Confidence stabiler. Das Variieren des Trainingsverhältnis zu gunsten von Protonen kann di Reinheit erhöhen.