A7)

1. Folgende Tabelle zeigt die verschiedenen Genauigkeitswerte auf den Datensätzen labor, yeast und car jeweils auf dem trainingsset und mit 10 fold cross-validation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | labor | | yeast | | car | |
| TS | 10CV | TS | 10CV | TS | 10CV |
| TP rate | 0,877 | 0,737 | 0,818 | 0,56 | 0,963 | 0,924 |
| FP rate | 0,089 | 0,28 | 0,062 | 0,145 | 0,02 | 0,056 |
| Precision | 0,896 | 0,748 | 0,821 | 0,55 | 0,964 | 0,924 |
| Instances | 57 | 57 | 1484 | 1484 | 1728 | 1728 |

1. Die folgenden drei Tabellen zeigen für die Datensätze labor, yeast und car die True-Positive-, False-Positive- und Precision-Rate jeweils für J48 mit Bagging und AdaBoost und RandomForest für verschiedene Iterationen. Wir beginnen jemeils bein einer Iteration und nehmen die aktuelle Anzahl Iteration im nächsten Schritt mal 2 bzw. 2,5. Somit ergeben sich die Iterationen 1, 2, 5, 10, 20, 50, 100, 200, 500 und 1000. Spätestens beim Schritt von 500 zu 1000 Iterationen verändern sich die Werte nur noch geringfügig, deshalb stoppen wir bei 1000. Diese asymtothische Annäherung an einen Wert beginnt bei den meisten Werten schon früher, teilweise bereits nach 20 Iterationen. Generell werden die Genauigkeiten mit mehr Iterationen präziser, allerdings auch aufwendiger. Daher sind 100 oder 200 Iterationen am sinnvollsten, da die erziehlten Werte schon sehr genau sind und der Aufwand noch akzeptabel.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Labor | Bagging mit J48 | | | AdaBoost mit J48 | | | RandomForest | | |
|  | TP rate | FP rate | Presicion | TP rate | FP rate | Presicion | TP rate | FP rate | Presicion |
| 1 | 0,719 | 0,358 | 0,611 | 0,737 | 0,28 | 0,748 | 0,842 | 0,223 | 0,841 |
| 2 | 0,789 | 0,298 | 0,786 | 0,772 | 0,284 | 0,77 | 0,807 | 0,242 | 0,805 |
| 5 | 0,825 | 0,256 | 0,823 | 0,825 | 0,21 | 0,825 | 0,877 | 0,181 | 0,878 |
| 10 | 0,86 | 0,214 | 0,861 | 0,895 | 0,126 | 0,895 | 0,877 | 0,181 | 0,878 |
| 20 | 0,842 | 0,2 | 0,841 | 0,895 | 0,126 | 0,895 | 0,895 | 0,149 | 0,894 |
| 50 | 0,895 | 0,149 | 0,894 | 0,877 | 0,135 | 0,879 | 0,895 | 0,149 | 0,894 |
| 100 | 0,877 | 0,158 | 0,876 | 0,86 | 0,168 | 0,86 | 0,895 | 0,149 | 0,894 |
| 200 | 0,877 | 0,158 | 0,876 | 0,842 | 0,2 | 0,841 | 0,895 | 0,149 | 0,894 |
| 500 | 0,86 | 0,191 | 0,858 | 0,877 | 0,135 | 0,879 | 0,895 | 0,149 | 0,894 |
| 1000 | 0,86 | 0,191 | 0,858 | 0,877 | 0,135 | 0,879 | 0,895 | 0,149 | 0,894 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Yeast | Bagging mit J48 | | | AdaBoost mit J48 | | | RandomForest | | |
|  | TP rate | FP rate | Presicion | TP rate | FP rate | Presicion | TP rate | FP rate | Presicion |
| 1 | 0,502 | 0,158 | 0,5 | 0,56 | 0,145 | 0,55 | 0,478 | 0,158 | 0,481 |
| 2 | 0,512 | 0,157 | 0,511 | 0,466 | 0,16 | 0,474 | 0,501 | 0,18 | 0,504 |
| 5 | 0,569 | 0,146 | 0,559 | 0,563 | 0,138 | 0,554 | 0,564 | 0,148 | 0,556 |
| 10 | 0,591 | 0,138 | 0,581 | 0,564 | 0,141 | 0,557 | 0,588 | 0,142 | 0,586 |
| 20 | 0,6 | 0,136 | 0,591 | 0,581 | 0,138 | 0,57 | 0,596 | 0,136 | 0,59 |
| 50 | 0,616 | 0,13 | 0,605 | 0,589 | 0,135 | 0,579 | 0,606 | 0,133 | 0,598 |
| 100 | 0,619 | 0,129 | 0,61 | 0,586 | 0,137 | 0,578 | 0,621 | 0,128 | 0,613 |
| 200 | 0,621 | 0,129 | 0,612 | 0,586 | 0,137 | 0,578 | 0,621 | 0,129 | 0,613 |
| 500 | 0,623 | 0,128 | 0,613 | 0,59 | 0,136 | 0,581 | 0,619 | 0,13 | 0,611 |
| 1000 | 0,621 | 0,129 | 0,612 |  |  |  | 0,621 | 0,129 | 0,613 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Car | Bagging mit J48 | | | AdaBoost mit J48 | | | RandomForest | | |
|  | TP rate | FP rate | Presicion | TP rate | FP rate | Presicion | TP rate | FP rate | Presicion |
| 1 | 0,909 | 0,054 | 0,91 | 0,924 | 0,056 | 0,924 | 0,836 | 0,16 | 0,836 |
| 2 | 0,916 | 0,064 | 0,916 | 0,924 | 0,056 | 0,924 | 0,875 | 0,159 | 0,866 |
| 5 | 0,928 | 0,044 | 0,93 | 0,955 | 0,026 | 0,955 | 0,914 | 0,072 | 0,914 |
| 10 | 0,931 | 0,045 | 0,933 | 0,961 | 0,016 | 0,962 | 0,936 | 0,046 | 0,936 |
| 20 | 0,935 | 0,037 | 0,937 | 0,964 | 0,016 | 0,964 | 0,946 | 0,032 | 0,947 |
| 50 | 0,941 | 0,029 | 0,943 | 0,966 | 0,01 | 0,967 | 0,945 | 0,026 | 0,947 |
| 100 | 0,939 | 0,033 | 0,94 | 0,967 | 0,01 | 0,968 | 0,945 | 0,026 | 0,946 |
| 200 | 0,938 | 0,03 | 0,94 | 0,971 | 0,008 | 0,971 | 0,946 | 0,027 | 0,947 |
| 500 | 0,936 | 0,032 | 0,939 | 0,972 | 0,006 | 0,973 | 0,949 | 0,023 | 0,951 |
| 1000 | 0,937 | 0,032 | 0,939 | 0,972 | 0,006 | 0,972 | 0,95 | 0,022 | 0,951 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | autos | | Abalone | | stick | |
| normal | diskretisiert | normal | diskretisiert | normal | diskretisiert |
| TP rate | 0,82 | 0,844 | 0,212 | 0,251 | 0,988 | 0,982 |
| FP rate | 0,046 | 0,046 | 0,097 | 0,105 | 0,11 | 0,236 |
| Precision | 0,833 | 0,844 | 0,201 | 0,232 | 0,988 | 0,981 |
| Größe | 69 | 165 | 2312 | 1231 | 61 | 46 |
| Blätter | 49 | 149 | 1183 | 1087 | 34 | 37 |

A8) Wir wählen die Datensätze autos, Abalone und stick aus

1. Die Daten sind auf den diskretisierten Daten generell ähnlich gut wie auf den normalen Daten
2. Auch die Baumgrößen werden manchmal besser, bei autos aber größer

A9)

Für die ersten Versuche verwenden wir nur einen Teil der Trainingsdaten, indem wir den Filter ... anwenden. Wir testen verschiedene Regelerner auf 5000 Trainingsdaten mit 10 fold cross –validation.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10 fold cross-validation with 5000 DP | Jrip | NaiveBayes | IBK | ConjunctiveRule | DecisionTable | OneR | PART | J48 |
| Correctly Classified | 74,66 | 75,64 | 89,48 | 19,76 | 56,34 | 19,18 | 76,36 | 73,14 |
| TP Rate | 0,747 | 0,756 | 0,895 | 0,198 | 0,563 | 0,192 | 0,764 | 0,731 |
| FP Rate | 0,028 | 0,027 | 0,012 | 0,095 | 0,049 | 0,092 | 0,026 | 0,03 |
| Precision | 0,749 | 0,775 | 0,897 | 0,077 | 0,558 | 0,154 | 0,763 | 0,731 |

Die besten fünf Lerner testen wir erneut mit 20 fold cross-validation auf 10000 Trainingsdaten.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 20 fold cross-validation with 10000 DP | Jrip | NaiveBayes | IBK | PART | J48 |
| Correctly Classified | 78,82 | 76,87 | 90,93 | 80,76 | 76,78 |
| TP Rate | 0,788 | 0,769 | 0,909 | 0,808 | 0,768 |
| FP Rate | 0,024 | 0,026 | 0,01 | 0,021 | 0,026 |
| Precision | 0,789 | 0,783 | 0,911 | 0,808 | 0,768 |

Das beste Ergebnis liefert IBK, also nearest neighbour. Deshalb testen wir für diesen Algorithmus verschiedene N (Anzahl betrachteter Nachbarn). Auch hier betrachten wir wieder 20 fold cross-validation auf 10000 Trainingsdaten. Die Werte sind generell gut mit einer Precision Rate von über 0,9. Die besten Werte werden mit N = 5 erzielt.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IBK-N Neighbours | N=1 | 2 | 4 | 5 | 6 | 7 | 9 | 10 | 100 |
| Correctly Classified | 90,93 | 91,39 | 93,84 | 94,25 | 94,01 | 94,04 | 93,85 | 93,76 | 87,9 |
| TP Rate | 0,909 | 0,914 | 0,938 | 0,943 | 0,94 | 0,94 | 0,939 | 0,938 | 0,879 |
| FP Rate | 0,01 | 0,01 | 0,007 | 0,007 | 0,007 | 0,007 | 0,007 | 0,007 | 0,014 |
| Precision | 0,911 | 0,918 | 0,94 | 0,944 | 0,942 | 0,942 | 0,94 | 0,94 | 0,893 |

Wir wollen die Testdaten mit IBK-5 klassifizieren. Unsere Vorhersage mit 10 fold cross-validation auf dem kompletten Trainingsdatensatz besagt:

|  |  |
| --- | --- |
|  | IBK - 5 |
| Correctly Classified | 96,598 |
| TP Rate | 0,966 |
| FP Rate | 0,004 |
| Precision | 0,966 |