DM Assignment 07

01.3.2016

Adithya Abraham Philip

1PI13CS008

Overview

The assignment involved experimenting with J48, Naive Bayes, JRip and Random Forest classification techniques.

Comparison

The classification techniques are arranged in order from best to worst (as perceived).

Reasoning:
1. Diabetes is a disease which becomes worse and potentially life threatening the longer it is left untreated. Therefore, it is encouraged to obtain **more true positives** at the **cost of having more false positives**.

2. This implies a **higher recall** is desired (even in the presence of **lower precision**), and therefore it is the primary measure on which they are ordered.

3. The measures used to sort the classifiers are, in order: **Recall, Precision, Accuracy, FP Rate.**

Notes: **F-Measure** is a function of recall and precision, **TP Rate** is same as recall, **Kappa** is a function of accuracy. These measures cannot therefore be used in tandem with the measures on which they depend.

Note: An **additional experiment** was performed on **Random Forests,** were besides setting **numFeatures**, the **effect of attributes being deleted was also examined**.

Notes: Attributes which increased accuracy on deletion were chosen to be removed first for the Random Forest initially, as determined by a combination of graphical analysis (which attribute had a more equal distribution of classes for each of its values was more irrelevant), and trial and error.

Note: Legend explaining notation for Random Forest is seen in Table 2.

Classifier	Accuracy	Kappa	TP Rate	FP Rate	Precision	Recall	F Measure
Naive	0.7630	0.4664	0.763	0.307	0.759	0.763	0.760
Random Forest [2]	0.7617	0.4649	0.762	0.306	0.758	0.762	0.759
JRip	0.7604	0.4538	0.760	0.322	0.755	0.760	0.755
Random Forest [1]	0.7500	0.4538	0.760	0.322	0.755	0.760	0.755
RF [n1]	0.7565	0.4414	0.757	0.333	0.750	0.757	0.750
RF [n3]	0.7552	0.4459	0.755	0.322	0.750	0.755	0.751
Random Forest [3]	0.7513	0.4396	0.751	0.322	0.747	0.751	0.748
Random Forest [0]	0.7487	0.4337	0.749	0.325	0.744	0.749	0.745
RF [n7]	0.7487	0.4327	0.749	0.327	0.744	0.749	0.745
RF [n5]	0.7435	0.4250	0.740	0.326	0.739	0.743	0.741
Random Forest [4]	0.7435	0.4199	0.743	0.335	0.738	0.743	0.739
J48	0.7383	0.4164	0.740	0.327	0.735	0.738	0.755
Random Forest [5]	0.7318	0.3907	0.732	0.353	0.725	0.732	0.727

Table 1. Comparison of various classifiers, ordered from best to worst

Notation	Effect on Features
Random Forest [0]	none removed, numFeatures not set
Random Forest [1]	skin removed
Random Forest [2]	skin, insulin removed
Random Forest [3]	skin, insulin, pressure removed
Random Forest [4]	skin, insulin, pressure, mass removed
Random Forest [5]	skin, insulin, pressure, mass, age removed
RF [n1]	numFeatures = 1

RF [n3]	numFeatures = 3
RF [n5]	numFeatures = 5
RF [n7]	numFeatures = 7

Table 2. Legend describing variations of attributes used with Random Forest

Observations

1. It may be **desireable** to pick a classifier with **lower accuracy**, as long as the **parameter** that is most **important** to the application at hand is **strong**. E.g. **Random Forest [1]** has lesser accuracy than **[3]** but is preferred because it has **higher recall**.
2. Although Recall and Precision tend to be inversely related in general, in this case, it is seen that they are positively correlated.
3. The accuracy of classification depends on the **attributes chosen** for classification. This is clearly visible with Random Forest, where deleting some attributes increases accuracy ([1], [2], [3]) (implies these attributes do not have a strong correlation to the class - diabetes), while continuing to delete attributes reduces accuracy (as relevant attributes begin to be removed).
   Most notably, Random Forest [2] (skin, insulin removed) is ranked second, while Random Forest [0] (no attributes deleted) is in the bottom half of the table.
4. It is observed that by **varying number of features** used in the Random Forest classifier: accuracy **initially decreases** with **increase** in number of features **(1->5)**, **then increases** with **increase** in number of features **(5->7).**
5. Naive Bayes classifier seems to give the best results, both in terms of our primary measure (recall) and accuracy.
6. Recall and TP Rate are the same (mathematically, not just empirically).