Assignment Three

Evaluation in Weka

Siobhán Grayson
12254530
6^{th} December 2013

1 Question One

For Question One, all evaluations were carried out in WEKA [1] using the Drexel-Stats dataset.

(a) IB1 Classifier Using 10-Fold Cross-Validation.

10-fold cross-validation is a common method of resampling, where the data is randomly split into 10 mutually exclusive subsets of approximately equal size. A learning algorithm is trained and tested 10 times; each time it is tested on one of the 10 folds and trained using the remaining 9 folds. The cross-validation estimate of accuracy is the overall number of correct classifications, divided by the number of examples in the data [2].

Doing this provides a more reliable estimate of the true accuracy of an algorithm [2]. Providing insight on how well a predictive model will generalize to an independent or unseen dataset as well as limiting the problem of overfitting that can arise when a model begins to memorize training data rather than learning to generalize from trend. With these procedures in place, the 1-NN classifier correctly classified 65.2% of instances using the *Drexel-Stats* data. Thus, the 1-NN classifier is estimated to perform reasonable well on unseen data.

(b) IB1 Using 10-Fold CV on Five Most Discriminating Features and Class Label

Feature subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. Feature selection prior to learning can be beneficial. Reducing the dimensionality of the data reduces the size of the hypothesis space and allows algorithms to operate faster and more effectively [3]. In order to determine the 5 most discriminating features, InfoGainAttributeEval in combination with Ranker were used.

InfoGainAttributeEval evaluates the worth of an attribute by measuring the information gain with respect to the class [4]. Ranker then ranks the attributes from highest to lowest according to their information gain score. The 5 most discriminating features that were selected from the *Drexel-Stats* dataset were: (1) Opponent, (38) Opp_Steals, (24) Opp_Field_Goal_Pct, (15) Def_Rebounds, and (28) Opp_Free_Throws_Made. Using just these 5 features and the class label, the performance of 1-NN was again assessed. This time Correctly Classified Instances was 91.3%. This is a very optimistic estimation, it is likely that the model has been overfitted to the dataset.

(c) Meta Filtered Classifier with Attribute Selection Integrated

Unlike the filter approach employed in 1.(b) above, attribute selection was integrated into the classification process using the structure outlined in the assignment documentation. This time 73.9% of instances were correctly classified. This is a relatively realistic generalization accuracy to achieve. It is not as high as 1.(b) but does still outperform 1.(a) where feature subset selection has not taken place. Illustrating that feature subset selection is indeed beneficial, removing redundant features and reducing noise [5].

(d) Comparison of the Above Assessments on the Impact of Feature Selection on Generalization Performance for IB1

In this assignment, 1.(c) gives a better assessment of the impact of feature selection on generalization performance for the 1-NN classification on the *Drexel-Strat* data. This is because 1.(b)'s assessment ignores the fact that the procedure has already "seen" the labels of the training data, and made use of them. Hence, an erroneously optimistic accuracy estimation is returned. In reality, it is impossible to say that the class labels and 5 features selected will be applicable or relevant to unknown independent data.

Instead, it should be treated as a form of training and included in the validation process as is the case with 1.(c). Here, a separate set of 5 features is formed in each fold of the cross-validation procedure. Test fold data is not used in features selection process as this data is held back. Measuring accuracy on a test set of examples is better than using the training set because examples in the test set have not been used to induce concept descriptions. Using the training set to measure accuracy will typically provide an optimistically biased estimate, especially if the learning algorithm overts the training data [2].

2 Question Two

For Question Two, all evaluations were carried out in WEKA using the ArtData dataset.

(a) Naïve Bayes Classifier Using 10-Fold Cross-Validation.

A naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with naive independence assumptions. It assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. This means that Naïve Bayes cannot handle continuous features well. This is apparent from the results obtained in this task, where only 34.5% of instances were correctly classified. It is clear that the continuous features in this data are being discretized as part of the classification.

(b) Naïve Bayes Using Using 10-Fold CV on Discretized Features.

Instead of continuous features being discretized as part of the classification, for this task a supervised attribute discretization preprocessing filter was applied to make it explicit prior to induction. The flags useBetterEncoding and useKonenko were set to true. Using this approach, the percent of correctly classified instances rose to 67.2%, almost double that of 2.(a). Thus, explicitly discretizing attributes achieves better generalisation accuracy than applying the classifier to non-discretized, scattered attributes.

(c) Is 2.(b)'s Estimate of Generalization Accuracy Realistic

For reasons similar to those outlined in section 1.(d) above, 2.(b)'s estimate of generalization accuracy is not realistic. The procedure has already "seen" the labels of the training data and uses this information during attribute discretization. This is not possible to implement on unknown independent data. Thus, 2.(b) produces an overly optimistic estimation of generalization accuracy.

One way of generating a more realistic estimate of generalization accuracy could be to implement a local rather than global supervised discretization method where discretization would be implemented during the induction process [2]. Using the same approach implemented in 1.(c), a meta filtered classifier was used incorporating the same supervised discretization as 2.(b) but this time as a part of the induction. As a result, 50.3% of instances were correctly classified, which is a more realistic estimate of generalization accuracy than that in 2.(b).

References

[1] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann P., & Witten I. H. (2009). *The WEKA Data Mining Software: An Update*. SIGKDD Explorations. Volume 11. Issue 1.

[2] Hall, M. A. (1999). *Correlation-based feature selection for machine learning* (Doctoral dissertation, The University of Waikato).

[3] Hall, M. A. (1999). Feature selection for discrete and numeric class machine learning.

[4] *Information Gain Attribute Evaluation OpenTox.* (n.d.). Retrieved from http://www.opentox.org/dev/documentation/components/infogainattribeval

[5] *Feature subset selection.* (n.d.). Retrieved from http://www.cse.unsw.edu.au/w̃aleed/phd/html/node164.html