Coursework B

Joseph Davidson - 071729468

1 Methods for Classifying Data with Missing Values

The purpose of this coursework was to learn some methods for dealing with missing data values in classification problems. As it turns out, there are multiple ways to deal with such situations, from something simple like ignoring the record to more advanced techniques such as predicting an answer, or range of answers for the missing data.

This short paper will quickly cover the following methods used for classifying such data:

- Ignoring the instance.
- Find the missing values.
- Imputation of the missing values.
- Using a specialised model (such as Max-Margin).

For reference, the first 3 items are referred to in [STP07], whilst the Max-Margin algorithm is the topic of [CHE+08].

Ignoring the instance: This method is the simplest and does exactly what it says on the the tin. The instance is ignored and often discarded from the set. It is often used when there is a penalty for incorrect classification – where it is safer to just refuse to classify the instance – or for data sets that have completely random missing data – so that no Imputation algorithm for the set is feasible.

Find the missing values: It may be possible to retrieve the missing values by some means such as from a 3^{rd} party source at evaluation time, or by performing a test to get the attributes. These actions typically incur a cost, but can complete the data set under scrutiny, leading to a better classification. Of course, this method only applies if there is a way to actually obtain the missing values.

Imputation: This refers to a class of methods that estimate the value or its distribution for a particular model. There are multiple types of imputation and a type should be picked for the situations that the classifier will likely find itself in. An example is by simply taking the mean or mode of the attribute over the rest of the set and imputing it.

Specialised models: There have been a handful of models developed that are robust to errors and omissions in the data to be classified. [CHE+08] goes into one of these – Max-Margin – in detail. The C4.5 algorithm uses a method of imputation to handle missing values in its input.

References

[CHE+08] Gal Chechik, Geremy Heitz, Gal Elidan, Pieter Abbeel, and Daphne Koller. Max-margin classification of data with absent features. *J. Mach. Learn. Res.*, 9:1–21, 2008.

[STP07] Maytal Saar-Tsechansky and Foster Provost. Handling missing values when applying classification models. *J. Mach. Learn. Res.*, 8:1623–1657, 2007.