

Ric Clarke
Director – Enabling Technologies
Australian Bureau of Statistics

Investigation Of A Machine Learning Approach For Automated Survey Coding In The ABS

Survey coding involves assigning a symbolic code from a predefined set of such codes to the answer given in response to an open-ended survey question. If a computer assigns codes without human interaction, then this is called automated coding (or autocoding). Manual coding, computer-assisted manual coding, and interactive coding all require some level of human interaction. The codes to which survey answers are assigned is determined by a classification, such as ANZSCO for Occupation, ANZSIC 06 for Industry, and ASCED for Education.

Existing autocoders in the ABS use a simple text retrieval technique that literally matches survey responses to the textual description of candidate codes. Unfortunately, the effectiveness of this approach is limited for complex classifications. In the case of the Occupation autocoder, the match rate on previous census data is less than 65%. Apart from text garbling, spelling errors and grammatical variation - which are a source of 'noise' for the autocoder - a major cause of match failure is associated with the limitations of dictionary-based syntactic analysis of unstructured text. There is often insufficient context to infer the intended meaning of an isolated survey response.

Enabling Technologies Section in the Technology Services Division (TSD) of the ABS has recently completed a brief research project to investigate an alternative approach to the autocoding problem where it is treated as an instance of multiclass text classification using supervised machine learning techniques. Here a software learning system inductively constructs a model of the association between survey responses and codes from a training set of pre-coded examples. Each response is considered a textual "document", and each code a "document category" to which these documents can be assigned. The learned model is a "classification function" that can be applied to the coding of new responses. In the project, we looked at a particular class of machine learning techniques - support vector machine (SVM) algorithms - which are very well suited to the classification of textual data, and can solve high dimensional problems with very few training examples.

The results of the initial investigation are very promising, and further work is planned to develop a prototype system for live testing in the 2011 Census. In this talk I will provide an overview of SVM machine learning theory, an outline of the research undertaken in the ABS, and the road map for future work in this area.