3 Memory-Based Reasoning Methods

Memory-based reasoning systems are a type of model, supporting the modeling phase of the data mining process. Their unique feature is that they are relatively machine driven, involving automatic classification of cases. It is a highly useful technique that can be applied to text data as well as traditional numeric data domains.

Memory-based reasoning is an empirical classification method. It operates by comparing new unclassified records with known examples and patterns. The case that most closely matches the new record is identified, using one of a number of different possible measures. This method can be applied to textual data as well as numerical data. That makes it very useful in comparing written documents. It has been used in a small number of cases, but has proven useful when conditions fit. Zurada and associates have been active in comparing memory-based reasoning against traditional data mining tools such as decision trees, neural networks, and regression for a variety of applications. They found memory-based reasoning to provide best overall classification when compared with the more traditional approaches in classifying jobs with respect to back disorders. This group found that traditional models did better in an application to predict debt re-payment in healthcare. Mixed results were obtained when testing fuzzy logic and memory-based reasoning to evaluate residential property values in real estate.

---

Matching

While matching algorithms are not normally found in standard data mining software, they are useful in many specific data mining applications. Fuzzy matching has been applied to discover patterns in the data relative to user expectations.\(^5\) Java software has been used to completely automate document matching.\(^6\) Matching can also be applied to pattern identification in geometric environments.\(^7\)

There are a series of measures that have been applied to implement memory-based reasoning. The simplest technique assigns a new observation to the preclassified example most similar to it. The Hamming distance metric identifies the nearest neighbor as the example from the training database with the highest number of matching fields (or lowest number of non-matching fields). Case-based reasoning is a well-known expert system approach that assigns new cases to the past case that is closest in some sense. Thus case-based reasoning can be viewed as a special case of the nearest neighbor technique.

Job applicant data: We use a small (appended) database to represent the training database of past job applicants, with ratings of success with the hiring firm. Some of these variables are quantitative and others are nominal. State, degree, and major are nominal. There is no information content intended by state or major. CA, NV, and OR are not expected to have a specific order prior to analysis, nor is major. (The analysis may conclude that there is a relationship between state, major, and outcome, however.) Degree is ordinal, in that MS and MBA are higher degrees than BS. However, as with state and major, the analysis may find a reverse relationship with outcome. So in the context of this example, degree is nominal.

Age and experience are continuous variables. For matching, these variables can be grouped into ranges, such as age under 25, 25–30, and over 30; or 0 experience, 1–2 years experience, and over 2 years experience. More is not necessarily better than less, however. The analysis will reach its conclusion one way or the other.