Clifton Phua, Kate Smith, Vincent Lee, Ross Gayler (not in publication order yet):
Chun Phua [Clifton.Phua@infotech.monash.edu.au]

****

On the empirical scoring of anomalous credit applications with pair-wise matching

Typical credit application fraud detection involves either known-fraud-matching using black-lists or supervised approaches using labeled data. Often, these labeled data approaches are operationally inefficient and ineffective, and do not take into account the complex nature of credit application data.

This paper introduces an unsupervised algorithm which defines the rational applicant behaviour and the normal social relationships between similar credit applications, and scores only the other similar, yet anomalous, credit applications.

Within each application, the identifiers are heuristically weighted to reflect their stable/changing nature. For example, names and driver license numbers have higher weights than addresses and telephone numbers. Across multiple applications, the identifiers of two applications are linked according to the chosen similarity metric(s). The algorithm determines if a pair of linked/similar applications is normal: submitted by the same applicant, and also by his/her family, housemates, colleagues, neighbours, and friends. For example, the applicant has changed his/her residential address, and applicant's family (first name, driver license number, and date of birth details are very different) has also applied for credit. If any pair of linked applications is considered anomalous, the algorithm then scores both applications by averaging the weights of the linked identifiers. A combined/final suspicion score for one application is obtained by adding scores from all the pair-wise matching with other applications.

There are other two contributions. First, to improve matching, significantly useful derived identifiers are appended to each application. A variety of phonetic string encoding, cultural naming conventions, and string frequency analysis are used. Second, to improve the approximate pair-wise matching of different identifiers, appropriate similarity metrics (distance measures) are utilised.

Results will be discussed for synthetic data and possibly real-world data too.