

## Business Applications of Artificial Intelligence with a Focus on Data Mining:

### A Literature Review

#### **Introduction**

As the field of artificial intelligence expands, as does its applicability to business environments. This applicability includes such areas as customer analysis and classification, improving efficiency, detecting fraud, and many others. Although each of these categories has something to gain from the use of data mining, we may find that privacy concerns should help regulate its usage. This paper will review available literature concerning each of the aforementioned topics in an attempt to gather and synthesize information which is available in scholarly journals for those who are interested or involved in the field of artificial intelligence.

#### **Background Overview**

Data mining is the process of analyzing data using such tools as theoretic models, lexicographic rules, and word association. Invented by IBM, the concept is gaining in popularity, especially in business settings where quickly processing large amounts of data is vital for company success. When combined with artificial intelligence, data mining is proving to be a successful component of business strategy.

#### **Customer Analysis and Classification**

One of the most published uses of data mining is the analysis of customer data for the purpose of classification. Once a customer is classified into one or more groups, businesses may direct targeted advertisements, mark them as potential risks or benefits

when considering service applications, or sell their information to other businesses for a profit.

Hodgkinson and Walker (2003) claim that the Credit Evaluation and Explanation Expert System (CEEES), written in Prolog, is capable of determining whether or not it would be in a business' best interest to approve an individual's credit line application. Such a decision is based upon the final analysis of a rating which is derived from a series of mathematical calculations. The authors explain:

The ratings are assigned using qualitative threshold rules that consider business credit history, asset size and liquidity, debt capacity, quality of management, market reputation, and position in market. The default probability is multiplied by the exposure at default and the loss given default to determine the expected loss. Unless the expected benefit is large enough to both subsume this expected loss and generate sufficient revenue for the financial institution, CEEES will recommend that the application for a credit line be rejected (Hodgkinson and Walker, 2003, 62).

The mathematical and logical structure of such calculations makes them predisposed for manipulation by electronic computers, and their “[resemblance to] the goal-directed backward chaining search used by Prolog” (65) leads them into the field of artificial intelligence.

Although Hodgkinson and Walker may be biased in favor of CEEES due to their affiliation with the project, they do present a logical and well backed argument as to why CEEES may be beneficial to the credit industry, how it works (including code excerpts), and where the project may be going in the future. The authors have not published any new information concerning CEEES since October, 2003; but the work which they have presented is sound and shows promise in aiding future systems.

## **Fraud Detection**

Ormerod, Morley, and Ball of Lancaster University in combination with Langley and Spenser of Logic Programming Associates (2003), correlate the use of ethnography and artificial intelligence for the purpose of detecting insurance fraud. They write that insurance fraud is a growing problem which, in the UK alone, cost the “industry \$1.4 billion [USD] in 2001” (650). In an attempt to gain back revenue by decreasing accepted fraudulent claims, the industry is beginning to turn to expert systems. The difficulty in this is that insurance fraud indicators tend to be dynamic in nature, while computer programs tend to be static. This means that machines must be taught to think by first becoming experts in the field, and then learn how to apply their information to a constantly changing environment (650). Although this was impossible just a few years ago, Hodkinson and Walker (2003) note that “[w]ith the improved computer hardware capabilities, large-scale expert systems have become feasible” (63).

Currently, expert systems are not being trusted as the sole judges of potentially fraudulent insurance claims, but they are being used as early detection mechanisms to mark and sort out suspicious insurance claims before they arrive at the hands of humans (Ormerod, *et al*, 2003, 651). Although no more recent information on this topic has been published, it may be interesting to see where this goes in the future.

## **Open Answers**

Yamanishi and Li (2002) detail another popular use of data mining. Customer feedback is an important aspect of business management, and expert systems are being used to mine open answer surveys with pleasant results. One such system, called the Survey Analyzer (owned by the NEC Corporation), has been used in Japan

(58) to analyze results from such industries as car manufacturing, Internet service providing, and beverage advertisement (62).

The system works by analyzing data using two rule subsets: classification rules, and association rules. Yamanishi and Li (2002) explain the distinctions between the two and claim that “[c]lassification rules are more suitable for summarizing open answers; [while] association rules are more suitable for discovering knowledge from open answers” (61). Although they have not released any proprietary code owned by NEC Corporation, they do present a convincing argument as to how each rule subset is used for the purpose of mining open answers in order to obtain interesting, and often surprising, correlations between two or more seemingly unrelated pieces of information.

### **Intrusion Detection**

According to Hu and Panda (2004), data mining has purposes beyond customer classification. Since data are only as reliable as they are authentic, preserving data integrity is of utmost importance. In order to accomplish this task, Hu and Panda (2004) propose a system which uses data mining to analyze a database’s usage log. From this log, expert systems may find trends such as when certain types of data are updated, viewed, or otherwise accessed, as well as any general trends concerning the order in which data is accessed. When data accesses which fall contrary to this trend are observed, database managers may be alerted to the possible intrusion and steps can be taken to quickly correct any maliciously altered data as well as control damage before it spreads (711 – 712).

As was the case with the previously mentioned authors, Hu and Panda may be biased towards their research, but they do present convincing mathematical arguments

and algorithms to back up their claims. Their article was published in March of 2004, and it may be interesting to read more about this when new information becomes available and the methods are potentially implemented.

### **Privacy Concerns**

Although data mining has numerous applications including the improvement of efficiency, of business and of database security; setbacks also exist, especially where the interest of personal privacy is concerned. Much information on this topic has been published within the last year, including a multi-national effort by Verykios and Theodoridis of the Academic and Research Computer Technology Institute in Athens, Greece; Bertino, Fovino, and Provenza of the Informational Science Department of the University of Milano in Milano, Italy; and Saygin of the Faculty of Engineering and Natural Sciences at Sabanci University in Turkey (2004).

The authors claim that privacy threats due to data mining are “well documented” (50) and that steps should be taken in order to avoid the exploitation of personal, sensitive data. By limiting access to raw data and rule sets, as well as removing personally identifiable information (such as names and addresses) from the data, one is able to protect individuals against those who would take harmful advantage of their information. Limiting access to raw data and rule sets is especially helpful, as it insures that participating data miners have access to only the results and the data which they contributed, and in no way can reconstruct the original raw data (e.g. more detailed information about specific individuals) by reversing the rule set algorithms (51).

As advances in data mining continue, the prospect for its misuse will increase and more information on the topic of data mining misuse is expected to be found.

According to Verykios, *et al*, “privacy preserving data mining is a novel research direction in data mining and statistical databases” (50). Once new minds are brought into the study, we may see an increase of privacy-friendly practices, which may, potentially, even be encouraged or unofficially required in the economic market if individuals demand to have their personal information protected rather than exploited.

## **Conclusion**

Although the combination of data mining with artificial intelligence is showing interesting results, we cannot ignore its potential for misuse. Current uses of data mining include customer analyses and classification, fraud detection, and the preservation of database integrity. Academic literature concerning the topic has been published in academic journals as recently as March 2004, while new and innovative information is promised as the field continues to mature.

## References

- Hodgkinson, L., & Walker, E. (2003). "An Expert System for Credit Evaluation and Explanation." *The Journal of Computing in Small Colleges*, 19(1), 62-72.
- Hu, Y & Panda, B. (2004). "A Data Mining Approach for Database Intrusion Detection." *ACM Symposium on Applied Computing*, 711-716.
- Ormerod, T., Morley, N., Ball, L., Langley, C., Spenser, C. (2003). "Using Ethnography To Design a Mass Detection Tool (MDT) For The Early Discover of Insurance Fraud." *CHI 2003: New Horizons*, 650-651.
- Verykios, V., Bertino, E., Fovino, I., Provenza, L., Saygin, Y., Theodoridis, Y. (2004). "State-of-the-art in Privacy Preserving Data Mining." *ACM SIGMOD Record*, 33(1), 50-57.
- Yamanishi, K., & Li, H. (2002). "Mining open answers in questionnaire data." *Intelligent Systems, IEEE [see also IEEE Expert]*, 17(5), 58-63.