Research Notes

* “Credit card fraud is a serious and growing problem. While predictive models for credit card fraud detection are in active use in practice, reported studies on the use of data mining approaches for credit card fraud detection are relatively few, possibly due to the lack of available data for research.” (602)
* Estimated $4 billion lost in 2008
* Supervised and unsupervised
  + “In supervised fraud detection methods, models are estimated based on the samples of fraudulent and legitimate transactions to classify new transactions as fraudulent or legitimate.” 602
  + “In unsupervised fraud detection, outliers or unusual transactions are identified as potential cases of fraudulent transactions.” 602
  + Both predict the probability of fraud in a given transaction, presumably by comparing it to the rest of them
* Predictive models are in active use
* Mail fraud when the credit card is intercepted before the customer could actually receive it (603)
* Phishing
* Logistical Regression, Support Vector Machines, and Random Forest
* Support vector machines are good at classifying
* Use statistics to detect fraud in journals (158)
* “The significant frauds that involved manipulation of financial statements and disclosures in the late 1990s and early part of this century gave added impetus to a fundamental shift in the conduct of audits.” (159)
* “Over the last several years, there has been an increased emphasis onthe detection of fraud as a key element of the financial statement audit” (159)
* “In 2000, the AICPA’s Public Oversight Board’s Panel on Audit Effectiveness pointed to a variety of necessary reforms to ensure the long-term viability of the audit. The significant frauds that involved manipulation of financial statements and disclosures in the late 1990s and early part of this century gave added impetus to a fundamental shift in the conduct of audits. A number of fraud schemes involved non-standard journal entries that were designed to make relatively simple adjustments between classes of accounts such that the financial statement results would show an improved position at the margin.” (159)
* Modern accounting information systems record transactions in a general ledger at the atomic ledger, and it is common for entities to have many thousands of entries in an accounting period
* “The effective and efficient data mining of journal entries requires comprehensive understanding of likely markers of fraudulent entries or adjustments, statistical properties of journal entries and the technological environment in which the client transacts the journal entries.” (164)
* “Fraud detection has become an increasingly important element of the financial statement audit.” (178)
* “Intrusion Detection Systems are security tools that provided to strengthen the security of communication and information systems. This approach is similar to other measures such as antivirus software, firewalls and access control schemes.” (708)
  + Common people, oxford comma!
  + In signature detection patterns are identified to provide a baseline, against which to compare activities
* Outliers are determined
  + Sensitivity would have to be determined somehow
  + Statistical significance?
* Credit card and identity theft
* “Anomalies are patterns in the data that do not conform to a well defined normal behavior. The cause of anomaly masy be a malicious activity or some kind of intrusion. This abnormal behavior found in the dataset is interesting to the analyst and this is the most important feature for anomaly detection.” (709)
* “Phua et al have done a detailed survey on various fraud detection techniques that has been carried out in the past few years. They have defined the professional fraudster, the main types and subtypes of known fraud, and also presented the nature of data evidence collected within affected industries.” (709)
* “Padhy et al provided a detailed survey of data mining applications and its feature scope. They stated that anomaly detection is an application of data mining where various data mining techniques can be applied” (709)
* “Clustering can be defined as a division of data into groups of similar objects. Each group, or cluster, consists of objects that are similar to one another and dissimilar to objects in other groups. Clustering algorithms are able to detect intrusions without prior knowledge.” (710)
* “Classification can be defined as a problem of identifying the category of new instances on the basis of a training set of data containing observations whose category membership is known. The category can be termed as class label. Various instances can belong to one or many of the class labels. In machine learning, classification is considered as an instance of supervised learning for example learning where a training set of correctly-identified observations is available. An algorithm that implements classification is known as a classifier. It is constructed to predict categorical labels or class label attribute. In case of anomaly detection it will classify the data generally into two categories namely normal and abnormal.” (710)
* “Infrequent pattern mining in data streams deals with extracting rare or unusual patterns from stream of data. Initially, pattern mining was focused on discovering frequent patterns, i.e., those patterns whose frequency of occurrence exceeds predefined minimum threshold.” (1998)
* “Recently, mining infrequent but useful patterns has gained the attention of data mining research community.” (1998)
* “In contrast to static dataset, mining useful patterns in data streams is a challenging task as it consists of an unbounded sequence of data which arrive to the system in a continuous manner. Windowing techniques such as sliding window, damped window, and landmark window have been proposed to capture the characteristics of evolving data streams.” (1998)
* “Mining frequent patterns finds application in market basket analysis, click stream analysis, web link analysis, genome analysis, etc. However, mining infrequent patterns are more important and relevant compared to frequent patterns in certain applications like network intrusion detection, credit card fraud detection, and anomaly detection.” (1998)
* Outliers vary considerably from other data
* Transaction Weighting Factor, Minimal Infrequent Pattern Deviation Factor, and Minimal Infrequent Pattern based Outlier Factor
* “Liu, Hsu, and Ma proposed MSapriori algorithm in order to address the problem of mining infrequent patterns that provide high confidence rules. The algorithm employes multiple minimum support thresholds to each item in the database. The minimum support of the rule is defined in terms of minimum support of the items that appear in the rule. Thus, the user is allowed to specify different support thresholds for different rules. However, the change in threshold values is driven by a subjective parameter, making the algorithm sensitive to user preferences. (1999)
* “Koh and Rountree (2005) developed Apriori-Inverse algorithm that defines both minimum and maximum support thresholds for generating rare items and discarding extremely rare items. The algorithm searches in level-wise bottom up fashion similar to Apriori algorithm. During each iteration, only those patterns whose support lies between minimum and maximum support are considered for further processing. The algorithm captures only the perfect rare items and fails to capture itemsets that are infrequent but frequent as individual items.” (1999)
* “Tsang, Koh, and Dobbie (2011) proposed a frequent pattern (FP) tree based algorithm for generating a set of rare items. According to the algorithm, entire database is scanned only once to find infrequent patterns whose support is less than minimum support. Authors have used information gain component to identify a set of rare association rules.” (1999)
* “information explosion” (134)
* “There is an urgent need for improved technology that can access, analyze, summarize, and interpret information intelligently and automatically.” (134)
* “discovering negative generalized knowledge from a relational database can reveal more interesting phenomena. For example, a medical database may tell us that although many patients are Taiwanese and numerous patients have H1N1, only a few Taiwanese have this disease.” (134)
* “Association rule mining is an important data mining approach that has been used to discover consumer purchasing behaviors from transaction databases.” (697)
* Two steps to association rule mining:
  + Generate all frequent itemsets
  + Generate all association rules that satisfy minimum confidence from already discovered frequent itemsets
* “However, relatively infrequent data as well as frequent data exist in the real world. While the previous association rule discovery techniques can be used to discover some rules based on frequency, they are insufficient to determine the importance of a rule composed of data items based only on their frequency.” (697)
  + Explores “Fuzzy specific rare itemsets” which are sets of items that rarely occur in the database together
* “Previous studies have tended to focus on the problem of discovering frequent patterns, meaning that only patterns that appear frequently in transaction data are mined. Patterns appearing in only a few data sets are not captured. In some cases, such as for the detection of computer virus attacks, fraudulent transactions in financial institutions or learning problems, those infrequent patterns, are more interesting than frequent patterns.” (697)
* “Rare itemsets are given by all itemsets that are not extracted by the standard frequent itemset generation algorithms, such as Apriori or FP growth.” (697)
* “The Apriori approach has been widely and successfully used to generate all frequent itemsets contained in a transaction database. Due to its great success and widespread usage, many variants of association rule mining have been proposed.” (698)
  + Three categories: Boolean, ordinal, and quantitative
* Relatively few algorithms to discover rare patterns
* “Szathmary et al proposed a method of discovering rare itemsets based on the Apriori algorithm which can be used to discover frequent itemsets. Briefly, the method can be divided into two steps: (1) all frequent itemsets and minimal rare itemsets are generated through the Apriori like algorithm MRG-Exp, (2) the mRIs discovered in the first algorithm are taken as seeds for the input data for the second rare itemset mining algorithm Arima.” (698)
* Rare patterns have many uses (699)
* “Data mining is a process of extracting valid, previously unknown, and ultimately comprehensible information from large datasets and using it to make crucial decisions.”(387)
* “Modern companies are awash in data on customers, clients, suppliers, and industry trends. But data is of little use without intelligence.” (387)
* “in the banking industry, data mining can be used in modeling and predicting credit fraud, bankruptcy, evaluating risk, performing trend analysis, analyzing profitability, and helping with direct marketing campaigns.” (387-8)
* Mostly a question of statistics
* “These methods usually make assumptions about data distribution, statistical distribution parameters, type or number of outliers.” These parameters cannot be easily determined (388)
* “Arning et al. Introduced a method for outlier detection which relies on the observation that after seeing a series of similar data an element disturbing the series is considered an outlier. Their method requires a function that can yield the degree to which a data element cauises the dissimilarity of the dataset to increase. It looks for the subset of data that leads to the greatest reduction in Kolmogorov complexity for the amount of data discarded.” (388)
* “Knorr and Ng presented the algorithms to detect distance-based outliers. They consider a data point O in a dataset T, a DB-outlier, if at least a fraction p of the data points in T lies greater than distance D from O. Their index-based algorithm executes a range search with radius D for each data point. If the number of data points in its D-neighborhood exceeds a threshold, the search stops and that data point is declared as a non-outlier, otherwise it is an outlier.” (388)
* “Data mining techniques are used to process huge amounts of information in order to extract hidden knowledge to be directly interpreted or exploited to feed other processes. In many cases, data mining techniques are used to discover patterns that can be of interest to a specific domain of application.” (1)
* “Rare patterns can be used in different domains such as biology, medicine and security, etc. For example, by analyzing clinical databases one can discover rare patterns that will help doctors to make decisions about the clinical care. In the security field, normal behavior is very frequent, whereas abnormal or suspicious behavior is less frequent.” (1-2)
* Defines a rare itemset as having a support greater than 0 and less than 3 of 5 transactions
* “proposes an approach to mine rare item-sets that is based on the Apriori algorithm used to mine frequent item-sets. The main idea consists at traversing item-set space by the Apriori algorithm used to mine frequent item-sets and collect at each level the item-sets that are usually pruned out in the original algorithm and are used as seed for a second algorithm in order to mine the remaining rare item-sets.” (4)
* “A minimal infrequent item-set is an infrequent item-set that do not have a subset of items which forms an infrequent item-set. In other words, an infrequent item-set is minimal if all its proper subsets are frequent.” (4)
* “Intrusion detection methods are of two types: anomaly detection and misuse detection. While anomaly detection techniques focus on the detection of user behavior that is considered as abnormal, signature detection focuses on the identification of a behavior that is similar to known cases that are considered as intrusions.” (5)
* Does Apriori in reverse
  + Start with the largest itemset, examine the subsets
  + Maximum support rather than minimum
  + Similar to candidate pruning
* “Mining association rules is a very important problem in the data mining field. It consists of identifying the frequent itemsets, and then forming conditional implication rules among them. This information is useful in improving the quality of many business decision-making processes, such as customer purchasing behavior analysis, cross-marketing and catalog design.” (5154)
* Gives a good definition of confidence
* “Discovering frequent itemsets is the computationally intensive step.” (5154)
* “Agrawal and Srikant (1994) proposed the Apriori algorithm to solve the problem of mining frequent itemsets. Apriori uses a candidate generation method, such that the frequent k-itemset in one iteration can be used to construct candidate (k + 1)-itemsets for the next iteration. Apriori terminates its process when no new candidate itemsets can be generated. DHP, proposed by Park et al. (1997), improves the performance of Apriori. It uses a hash table to filter the infrequent candidate 2-itemsets and employs database trimming to lower the costs of database scanning. However, the aforementioned methods cannot avoid scanning the database many times to verify frequent itemsets”
* “Unlike Apriori, the FP-growth information of the transaction database. Without candidate generation, FP-growth uses a recursive divide-and-conquer method and the database projection approach to find the frequent itemsets” (5154-5155)
* “In this paper, we propose the IFP-growth (Improved FP-growth) algorithm to improve the performance of FP-growth. First, the IFPgrowth employs an address-table structure to lower the complexity of mapping frequent 1-itemsets in an FP-tree. Second, it uses a hybrid FP-tree mining method to reduce the need for rebuilding conditional FP-trees. Memory space can be saved and the cost of re-constructing conditional FP-trees can be reduced. We also present experimental results, and compare our methods to several existing algorithms, including FP-growth and nonordfp. Simulation results show that IFP-growth mines frequent itemsets efficiently with less memory space requirement. Under various minimum supports, IFP-growth can outperform FP-growth and nonordfp in execution time”
* Turned out to be far faster than the FP-growth algorithm
* “Itemset mining is an exploratory data mining technique widely used for discovering valuable correlations among data. The first attempt to perform itemset mining was focused on discovering frequent itemsets, i.e., patterns whose observed frequency of occurrence in the source data (the support) is above a given threshold. Frequent itemsets find application in a number of real-life contexts. However, many traditional approaches ignore the influence/interest of each item/transaction within the analyzed data. To allow treating items/transactions differently based on their relevance in the frequent itemset mining process, the notion of weighted itemset has also been introduced. A weight is associated with each data item and characterizes its local significance within each transaction.”
* “In recent years, the attention of the research community has also been focused on the infrequent itemset mining problem, i.e., discovering itemsets whose frequency of occurrence in the analyzed data is less than or equal to a maximum threshold.”
  + Good for me
  + Contradicts what I’ve said previously, but that’s okay by me
* “Infrequent itemset discovery is applicable to data coming from different real-life application contexts such as (i) statistical disclosure risk assessment from census data and (ii) fraud detection [7], [8], [9]. However, traditional infrequent itemset mining algorithms still suffer from their inability to take local item interestingness into account during the mining phase”
  + Touched on this with closed itemsets
* “As future work, we plan to integrate the proposed approach in an advanced decision-making system that supports domain expert’s targeted actions based on the characteristics of the discovered IWIs. Furthermore, the application of different aggregation functions besides minimum and maximum will be studied”
* Need to spend some more time on this one, seems similar to what I’m proposing to do

# Bibliography

|  |  |
| --- | --- |
| [1] | S. Bhattacharyya, S. Jha, K. Tharakunnel and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Elsevier,* vol. 50, no. 3, pp. 602-613, February 2011. |
| [2] | R. S. Debreceny and G. L. Gray, "Data mining journal entries for fraud detection: An exploratory study," *International Journal of Accounting Information Systems,* vol. 11, no. 3, pp. 157-181, September 2010. |
| [3] | S. Agrawal and J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques," *Procedia Computer Science,* vol. 60, pp. 708-713, 2015. |
| [4] | C. S. Hemalatha, V. Vaidehi and R. Lakshmi, "Minimal infrequent pattern based approach for mining outliers in data streams," *Expert Systems with Applications,* vol. 42, no. 4, pp. 1998-2012, March 2015. |
| [5] | Y.-Y. Wu, Y.-L. Chem and R.-I. Chang, "Mining negative generalized knowledge from relational databases," *Knowledge-Based Systems,* vol. 24, no. 1, pp. 134-145, February 2011. |
| [6] | C.-H. Weng, "Mining fuzzy specific rare itemsets for education data," *Knowledge-Based Systems,* vol. 24, no. 5, pp. 697-708, July 2011. |
| [7] | D. Yu, G. Sheikholeslami and A. Zhang, "Finding Outliers in Very Large Datasets," *Knowledge and Information Systems,* vol. 4, pp. 387-412, 2002. |
| [8] | M. Adda, L. Wu, S. White and Y. Feng, "Pattern Detection with Rare Itemset Mining," *International Journal on Soft Computing, Artificial Intelligence and Applications,* vol. 1, no. 1, pp. 1-17, August 2012. |
| [9] | K.-C. Lin, I.-E. Liao and Z.-S. Chen, "An improved frequent pattern growth method for mininig association rules," *Expert Systems with Applications,* vol. 38, no. 5, pp. 5154-5161, May 2011. |
| [10] | L. Cagliero and P. Garza, "Infrequent Weighted Itemset Mining using Frequent Pattern Growth," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, no. 4, pp. 903-915, April 2014. |