Mine Your Business—A Novel Application of Association Rules for Insurance Claims Analytics

Lucas Lau and Arun Tripathi, Ph.D.

Abstract: This paper describes how a data mining technique known as Association Rules can be applied in the analysis of insurance data to gain useful and actionable business insights. The technique is illustrated via its application to Workers Compensation (WC) insurance data. This case study shows how the Association Rules technique can be used to potentially reduce claim costs, manage claims, and help prevent injuries at workplace. While this case study uses WC data for the case study, the analytic technique presented in this paper can be extended to other types of insurance data.

Keywords: association rule; workers compensation; Apriori; confidence; lift; support

1. INTRODUCTION

The Pareto principle, popularly known as the 80/20 rule, states that for many events, 80% of the effects come from 20% of the causes [Wikipedia]. Workers Compensation insurance is not an exception to this rule – indeed the rule applies to Workers Compensation business in a more exaggerated form. For example, 60% of Workers Compensation claim dollars arise from 5% of the claims for some self-insurers. Common examples of such claims are lower back injury and eye injury caused by foreign objects. In addition to the cost to the insurer, these injuries obviously have an adverse impact on the life and health of the person affected and also result in lost productivity for the employer. Even a relatively small number of such high-loss claims can result in a disproportionately high cost to the insurer.

Clearly, any methodology that provides insights into the characteristics of such injuries and resulting claims will be immensely helpful in the following ways:

- First, the insights into the nature of the injuries and risk factors that the workers are exposed to can be used to create an evidence-based and disciplined injury prevention program at workplace and to enhance any existing safety programs already in place. This will help prevent the injuries from occurring in the first place, which is beneficial for the workforce as well as insurers. As a simplified example, if an employer is seeing a lot of eye injuries at the workplace, they can prevent them by creating a safety program that requires employees to wear goggles and other protective gear.

- Second, understanding the correlation between the nature of the injuries and the size of the

resulting claim can be used for a more effective claims management, providing significant potential cost savings to the insurer. The claims department of any insurer has limited resources, which can be utilized in a far more optimal manner if there is a way to predict the size of the claim as soon as the claim is initiated.

In this paper, we explore how a data mining technique known as Association Rules can be used to achieve the above goals.

2. ASSOCIATION RULES

Association Rules is a data mining technique commonly used in retail business to understand the purchase behavior of the consumer – what groups of items do consumers tend to purchase together? As an obvious example, such analysis might reveal that people tend to buy shampoo and conditioner together. This information can then be used to create appropriate sales promotions or the placement of the products in a supermarket.

How can this technique help a Workers Compensation insurance company? The answer is: in the same manner as it has helped the retail industry, by finding the types of items that occur together. The idea is to apply this type of analysis to historical Workers Compensation claims data to find out associated or co-occurring features of the claim. Such an analysis might reveal, for example, that a large fraction of leg injuries result in fractures, which also result in large claims. This insight can then be used to improve the safety program to take preventive measures – e.g. leg safety gear, better training etc.

2.1 A Simplified Numerical Illustration

Before giving a concrete example of how Association Rules can be applied in WC insurance, it is useful to review some basic concepts associated with this technique. These concepts are essential to understand and interpret the results of such an analysis. To that end, we focus on a toy example of customer purchases, as shown in the table below.

Customer	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Table 1: Supermarket Data

The above table contains hypothetical supermarket data extracted from Wikipedia. Each row represents a transaction per customer. Columns are the purchased food items. For example, the first row tells us that customer 1 bought milk and bread, but not butter and beer, and so on.

Now, suppose the goal is to find out the probability of customers buying butter, if they also brought bread and milk. This statement (the goal or rule) can be represented in the form:

{Bread, Milk} => {Butter}.

Before we continue with the illustration, we need to define a few definitions. In the above rule:

{*Bread and Milk*} is called ***antecedent*** (left-hand-side or LHS, the "if" part of the rule).

Butter is called ***consequent*** (right-hand-side or RHS, the "then" part of the rule).

The reason for translating the goal statement in the above form is because this is the format in which the Association Rules are produced at the end of the analysis.

How to interpret the rules?

In our toy example, 20% (1 out of 5) of customers bought milk, bread and butter together. This is the unconditional probability of buying milk and bread and probability of buying butter together. This quantity is also known as the ***Support*** for the itemset {Bread, Milk, Butter}. Support indicates how frequently certain combinations of items in both antecedent and consequent occur together in the data.

Of the two customers that bought both milk and bread, one of them also bought butter. This means 50% of the customers who bought both milk and bread, also bought butter. This conditional probability is called the ***Confidence***. It is a measure of how often the consequent is true, given that the antecedent is also true. A large value of Confidence means the rule is pointing to a strong association between the antecedent ("if" part of the rule) and the consequent ("then" part of the rule).

Lift is a measure of how predictive the rule is, compared to random association. It is defined as the ratio of the observed confidence to expected confidence. When lift is greater than 1, then the resulting rule is better at predicting the result than a random guess. In our example the lift is calculated to be: $Pr\{Bread, Milk\ AND\ Butter\}/(Pr\{Bread, Milk\} \times Pr\{Butter\}) = 0.2/(0.4 \times 0.4) = 1.25$.

3. ANALYSIS FRAMEWORK

The general framework of developing Association Rules has four steps: 1) Define your business goal, 2) Prepare data, 3) Generate rules in rules' engine and 4) Interpret and implement rules. Sample Workers Compensation insurance data will be used to illustrate.

3.1 Define the Right Goal(s) to Get the Best Out of the Rules

Typically, an Association Rule analysis will result in many rules, containing different itemsets. In order to select rules that are relevant and useful, we should have a clear picture of which item(s) we are looking for. Then we can only focus on the rules that contain those item(s). The examples below illustrate how we might go about setting the business goals.

Sample Goal #1: Find the major causes of eye injuries in the claim data.

In this case, we should look for all rules that have "Eye" as consequent. By doing that, safety manager can find out what are the common factors and circumstances associated with eye injuries. Rules resulting from such an analysis may look like the following:

{Driver, Foreign Object} => ***{Eye}***

{Day Shift, Acidic Bottle, Foreign Object} => ***{Eye}***

Sample Goal #2: What are the major injury types associated with a focused group, e.g. day-shift driver?

In this case, we should look for all rules that have "day shift, driver" in the antecedent. For example:

{***Day Shift, Driver***, *Obstacle, trip*} => {*Sprain*}

{*Cut, Glass,* ***Day Shift, Driver***} => {*Laceration*}

3.2 Prepare the Data

Depth and breadth of data is essential for a meaningful analysis. The claim data should capture details of the incident and injury such as time, place, cause of injury, affected body part, location, etc. The level of data for input depends on the specification of statistical package or software. In most cases, data is organized in transactional form (one row per transaction or claim detail) or at claim level.

3.3 Generate Rules in Rules Engine

A number of statistical packages available in the market have the capacity to perform Association Rule analysis. Behind the scenes, a complex algorithm is used to complete the sophisticated calculations. Apriori is the most popular algorithm used by statistical packages. The algorithm uses a "bottom-up" approach which starts by calculating the frequency of occurrence of 1-itemset and extends it to *n*-itemset until no further extension is found. The process is illustrated in the following figure.

Below is a hypothetical closed injury claim data. Each column represents a cause of injury. The minimum frequency threshold is three in this case. Apriori algorithm summarizes the frequency of 1-itemset {Foreign Object, Driver, Debris, Eye} first. The frequency distribution is the left most table underneath of the data table. Next, the algorithm moves on to 2-itemset. The frequency distribution is the middle table below data table. The 2-item set {Foreign Object, Debris} has a frequency of 2, which is below the minimum threshold. It is excluded from the analysis. Then the algorithm continues counting the frequency for 3-itemset. The frequency of the 4-itemset is excluded because it falls below minimum threshold. The counting stops at 4-itemset as it exhaust all the possible combination in the data.

APRIORI CALCULATION EXAMPLE

Claim	Foreign Object	Driver	Debris	Eye
1	1	1	1	1
2	0	1	1	1
3	0	1	1	0
4	1	1	0	1
5	1	1	1	1
6	0	1	0	1

Table 2: Claims Data for Apriori Algorithm Illustration

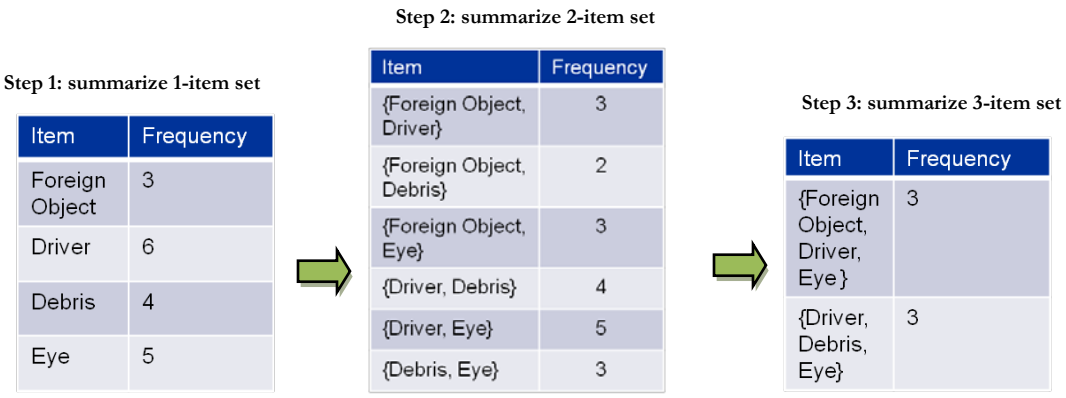

Figure 1: The flow of Apriori Algorithm

3.4 Interpret and Implement the Rule

The process of getting the Association Rule analysis to yield rules that are at once interesting (not obvious or trivial) and actionable requires both art and science. First, during the data analysis phase, one must experiment with the minimum support and confidence levels for the analysis. If the level of support and confidence is set to smaller values during the analysis, a large number of rules will emerge, which one would need to sift through. On the other hand, if the level of support and confidence is set to larger values, fewer rules will come out of the analysis, but some potentially interesting rules about the combinations that don't occur too frequently will drop out. If possible, relatively small values of support and confidence should be specified during the analysis, and then the results should be examined visually to look for interesting rules.

Even after the minimum thresholds for support and confidence have been optimized, the resulting rules must be carefully examined to get meaningful results. For example, many high-confidence rules will be obvious or trivial, providing no new insight. For example, one may find a rule such as: {Cut, Puncture, ScrapebyBrokenGlass} => {Lacerations/Cuts}. While such a rule might have a high degree of confidence, it is a rule that is trivially true, since both LHS and RHS of the rule point to Cuts.

Further, some rules may be interesting but may not be actionable due to legal and other business constraints. For example, the rule may involve protected characteristics (e.g. Gender, Race etc.)

However, a close examination of all the resulting rules may reveal some interesting and useful patterns, which are not only statistically significant but also actionable with a clear tangible benefit to the business. The following case study illustrates the points discussed above.

4. APPLICATION OF ASSOCIATION RULE ANALYSIS IN WORKERS COMPENSATION INSURANCE

In Workers Compensations insurance, understanding the cause of injury is crucial for preventing repeat injuries. This section describes how Association Rule analysis can be used to reduce both claim frequency and severity. The technique enables insurers and self-insurers to understand the pattern of circumstances related to the injuries.

For example, a waste management company incurs eye injury claims costing $500,000 annually. These injuries and the associated insurance costs could potentially be lessened by enforcing an effective safety program in the workplace. However, for the safety program to be effective, the

safety manager needs to understand the circumstances (what/where/when) associated with the eye injuries. Association Rule analysis is very well suited for this task.

The table below shows some sample results after running the Association Rule analysis on historical claims data from this company. The rules were built for two data sets – (a) all claims, and (b) Just the top 3.5% claims, based on the claim amount, which account for 60% of the total claim cost.

Rule	Data Set	Antecedent	Consequent	Support	Confidence	Lift
1	All Claims	{Day Shift,ForeignBody}	{Eye(s)}	1.55%	0.81	16.57
2	Top 3.5% Claims	{DRIVER,Day Shift,Lowerleg(s)}	{Fracture}	1.16%	0.80	9.17
3	Top 3.5% Claims	{Day Shift,StrainorInjuredbyLifting,Tuesday}	{Strain}	1.74%	1.00	2.23
4	Top 3.5% Claims	{Night Shift}	{Monday}	2.91%	0.40	1.68

Table 3: Sample Association Rule Result

These results provide insights into the features associated with different types of injuries. In addition, as we will see, they also illustrate why a manual inspection of the rules is needed to select useful and actionable rules.

The first rule in this table tells us that eye injuries are highly associated with day shift, and the hazard of being hit by a foreign body. Of all the claims associated with day shift and injury by a foreign body, 81% result in eye injury. This result is highly significant, with a lift of over 16. This means that workers are 16 times more likely to sustain an eye injury by a foreign body during the day shift, compared to what can be expected if the eye injuries were not correlated with {Day Shift, ForeignBody}. This suggests that the company will benefit from implementing and improving the eye safety program for its employees.

The next rule is related to the fracture claims. This rule tells us that 80% of the lower leg injuries for day-shift drivers result in fractures. Again, this rule is very significant, with a lift of over 9. In addition, the analysis of claim data shows that fracture claims are some of the costliest on a per-claim basis. This suggests that the company would benefit from implementing safety programs to prevent lower-leg injuries.

However, as mentioned above, a typical Association Rule analysis will also yield rules that are either trivial or non-actionable. For example, the third rule in the table above tells us that the day-shift workers suffering strain by lifting on Tuesdays are 100% associated with strain. This is trivially true: strain implies strain!

The final rule tells us that 40% night-shift injuries occur on Mondays. On the face of it, this rule does not provide a clear explanation of what is going on here and does not suggest a course of action. Knowledge of the business may reveal the reason behind this pattern. Perhaps a large number of night-shifts occur on Mondays, after the weekend parties at both businesses and homes. So perhaps this rule is just reflecting this pattern of how the night-shifts are distributed during the week.

These examples illustrate that Association Rule analysis can provide interesting insights about the activities and items associated with a given type of injury. These insights can then be used by the safety manager to design an effective and focused program by offering preventive measures and protective equipment to its employees who are at the highest risk of being injured.

5. CONCLUSION

In this article, we have illustrated how a data mining technique known as Association Rules, which is commonly used in retail business, can be successfully applied in the analysis of insurance claim data to gain interesting and useful insights about the circumstances surrounding the claims. These insights can be used to minimize future incidents leading to such claims. While the case study shown here is based on Workers Compensation claims, the technique can easily be applied to other lines of business as well. Even if other types of analytic techniques are already being used to analyze claims data, this technique can provide valuable complementary value.

A Note Regarding Software

All association rules analyses discussed in this paper were processed using the open source R statistical computing package. R is available at http://ww.r-project.org. Once the base R package has been installed, the association package "arules" is required in order to process association analysis.

Acknowledgments

We would like to thank James Guszcza for his guidance in editing the draft and also for introducing us to the articles by Michael Hahsler. We also thank Peter Wu, Frank Zizzamia and David Duden for their constructive and informative comments on an earlier draft of this paper.

5. REFERENCES

R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 2003.

Berry, M. J. A., Linoff, G. S., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2003, John Wiley & Sons.

M. Hahsler, B. Grun, K. Hornik, and C. Buchta, "Introduction to arules – A computational environment for mining

association rules and frequent item sets," The Comprehensive R Archive Network, March 2009.
M Hahsler, B Grun and K Hornik, "arules – A computational environment for mining association rules and frequect item sets," *Journal of Statistical Software,* 2005, Vol XVI, 1-25.
Hastie, T., Tibshirani, R., Friedman, J. H., The Elements of Statistical Learning, 2003, Springer.
Surjandari dan Annury Citra Seruni, Isti, "Design of Product Placement Layout in Retail Shop using Market Basket Analysis," *Makara Teknologi* , Vol. 9, No. 2, 43-47.

Abbreviations and notations

Collect here in alphabetical order all abbreviations and notations used in the paper

APD, automobile physical damage
CL, chain ladder
DFA, dynamic financial analysis
GLM, generalized linear models
OLS, ordinary least squares
ERM, enterprise risk management

Biographies of the Authors

Lucas Lau is Senior Consultant at Deloitte Consulting LLP in Advanced Analytics and Modeling Practice. He is responsible for business analytics, statistical modeling and business implementation for both Property & Casualty Insurance and Life Insurance. Mr. Lau has a master degree in Statistics from Columbia University.

Arun Tripathi is a Senior Consultant at Deloitte Consulting LLP in Advanced Analytics and Modeling Practice. He has over 20 years of predictive modeling experience in diverse areas such as High Energy Physics, Property & Casualty Insurance, Life Insurance, Retail etc. Dr. Tripathi has a Ph.D. degree in Physics from The Ohio State University.