Characterizing Problem Gamblers in New Zealand: A Novel Expression of Process Cubes

Suriadi Suriadi, Teo Susnjak, Agate M. Ponder-Sutton, Paul A. Watters, Christoph Schumacher

Massey University, Albany, Auckland 0632, New Zealand
s.suriadi@massey.ac.nz

Abstract. This paper reports on the challenges and lessons learned from our case study which uniquely integrates a mixture of process mining, data mining, and confirmatory statistical techniques to explore and characterize the variations in gambling behaviours exhibited by gamblers in New Zealand. We demonstrate how we weaved techniques from these three disciplines to understand the variety of behaviours exhibited by gamblers, and to provide assurances of the correctness of our results. This case study also demonstrates how such a combination of techniques provides a rich set of tools to undertake an exploratory data analysis project that is guided by the process cube concept.

Keywords: data mining, process mining, statistics, problem gamblers

1 Introduction

This paper reports on the challenges and lessons learned from our case study where a mixture of process mining, data mining, and confirmatory statistical techniques are applied to analyse a data set containing all bets recorded by a gambling service provider in New Zealand (NZ). This study attempts to identify and characterize various groups of gamblers (with a focus on problem gamblers) directly from the data. The nature of this case study is exploratory: we do not 'label' our data with various classes of gamblers; rather, we attempt to learn how many groups of gamblers can be discerned from the data.

The exploratory nature of this study calls for the use of unsupervised learning techniques, such as k-means clustering [4] as the starting point of analysis. However, clustering analysis alone is not sufficient as it does not offer any insights into the behaviour of gamblers seen in each cluster. Our case study *demonstrates how both process mining analysis*, with its ability to extract detailed insights from fine-grained and chronologically-arranged data [8, 11], and *confirmatory statistics*, can be strategically weaved together with clustering analysis to not only understand the variety of behaviours exhibited by gamblers, but also to evaluate our results. Furthermore, we highlight how the combination of these techniques can be used as an expression of the process cube concept [14].

The main contribution of this paper is on the reporting of challenges and lessons learned in the application of these three classes of analysis techniques,

In: S. España, M. Ivanović, M. Savić (eds.): Proceedings of the CAiSE'16 Forum at the 28th International Conference on Advanced Information Systems Engineering, Ljubljana, Slovenia, 13-17.6.2016, published at http://ceur-ws.org

highlighting practical challenges that arise in analysing a relatively large size of data with limited computational resources. Section 2 summarizes the approach taken in this case study. Section 3 details the analyses performed at each stage of the case study, along with the challenges and lessons learned. A discussion about the related work is provided in Section 5, followed by the conclusion.

2 Approach

While we applied a range of techniques from multiple disciplines, our case study employed the PM^2 [16] approach because: (1) the starting point of our analysis was an event log (the log used in our study consisted of betting *events* per account holder), of which process mining is designed for, and (2) the PM^2 methodology is rather detailed in its guidance on each stage, and flexible enough to allow the inclusion of other types of classical data mining techniques (e.g. clustering).

We also applied the concept of *process cubes* - a concept born from the Online Analytical Processing (OLAP) domain. Using this approach, event logs are organised into various cells based on multiple dimensions [14] and these cells can then be merged or further split using typical OLAP operations, such as slice, dice, roll-up, and drill-down. However, as suggested by van der Aalst [14], the nature of event logs are such that special considerations are needed to perform those operations. This paper shows some of the techniques that can be applied to achieve those operations.

3 Case Study

This case study focuses on the data processing, data mining, clustering and analysis, and evaluation stages of the PM^2 methodology. The questions that the stakeholder in our case study (an economist) wanted to address were: (RQ1) how many groups of gamblers can be discerned from the data set, and (RQ2) which users were problem gamblers from the data?

The data set was extracted from a NZ gambling provider. It contained information about all the bets placed in New Zealand from August 2006 to May 2014. The data set extracted was in event log format where each line of data represents a betting event. Given the size of the raw data (80 GB in compressed text format) and the limited computational resources at our disposal (three 8-16GB/i5 Core workstations and one 32GB/i5 Core virtual machine), we focused our analysis on gambling data from a 9-month window (Aug 2013 to May 2014).

Data Processing. The dataset consisted of 11,311,892 betting events executed by 91,405 account-holders. The data was pre-processed because (1) our experiences showed that existing implementations of many process mining techniques do not scale well; many process mining case studies used event logs that are significantly smaller (between a few thousand events to over 1 million events, e.g. [2]); (2) the raw data is not in an event log format consumable by process mining analyses.

Applying the concept of process cube [14] calls for slicing and dicing event logs from multiple dimensions. As our case study is exploratory (without any clear filtering dimensions), we applied an unsupervised k-means clustering [4] technique to find the 'natural separation' of gamblers in the data set. Three problem gambling features from the data were used as clustering criteria: bet frequency, the ratio (in dollar) of winnnings over amount lost (*win/loss dollar ratio*, and ratio of the number (count) of winning over lost bets (*win/loss ratio*). Next, an event log was derived for each cluster using a gambler's account identifier as the case ID in the event log.For the activity field, the data was preprocessed to approximate a gambler's *clawback* behaviour: an activity name was created as the concatenation of (1) the outcome of the previous bet (`win` or `loss`) and (2) the amount of money placed in the next immediate bet as a proportion of the amount of money placed in the previous bet: less than or equal to the previous bet amount (<= 1), up to double the previous amount (`1to2`), or more than double (>`2`). For example, `loss_`<= 1 would mean that a user having lost the previous bet, had subsequently placed another bet where the amount bet was less than or equal to the previous bet. Through data pre-processing, we obtained 7 manageable-size event logs for each cluster for deeper analysis. Each cluster is *colour-coded for referencing purposes.*

Challenges. The large dataset was problematic given the lack of scalability of the software tools used. Attempts to generate an XES file (i.e. the log format expected by most process mining tools) presented challenges: we were limited by the number of events that could be imported in Disco (`www.fluxicon.com`) tool; we could have used the ProM Tool (`www.prcessmining.org`) to convert the original CSV-formatted log into XES; however, the generated XES file would have been substantially larger than the original CSV data, thus would not have scaled well.Trace clustering algorithms, e.g. [17], could not have been applied as they would have required the entire event log to be analysed (not feasible with our resources). We addressed this by using k-means clustering [4] based on aggregated features at case-level granularity. This required deciding the optimal number of clusters (k) to be used. We combined the advice from the stakeholder with the *Within Sum of Squares* (WSS) analysis which showed that the cohesiveness of the clusters converged optimally in the range of k=7 and k=8. We therefore decided to use 7 clusters.

Mining and Analysis. Our analysis focused on extracting key markers for problem gamblers across clusters using process mining techniques. Clawback behaviour is a key problem gambling psychological feature. We studied the clawback behaviours by generating *process models* for each cluster using the Disco tool. Then, through visual comparisons (similar to Partington et al. [8] and Suriadi et al. [12]), we extracted 9 distinct flow patterns in each cluster.

To assert statistical significance w.r.t the differences in the distribution (in terms of frequency) of these 9 patterns across all clusters, Kruskal-Wallis tests in conjunction with Dunn's tests were performed. These tests showed that *the distribution of these patterns across the seven clusters were indeed significantly different* at the p-value of 0.05. In addition, from box-and-whisker plots, we also observed that the distribution of these 9 patterns were quite distinct in the blue and black clusters (see Figure 2, left diagram, for example).

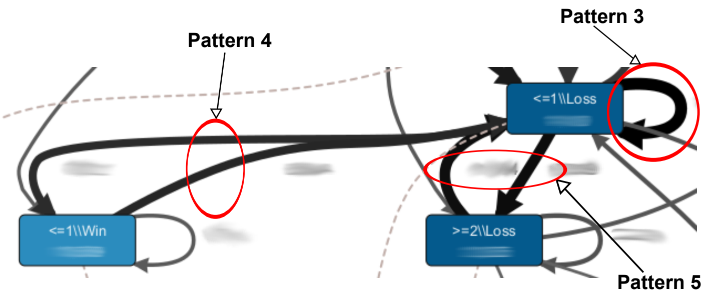

Fig. 1: A visualisation of clawback behaviours identified from the process models of each cluster. Pattern 3 depicts a gambler who repeatedly lost bets and kept on betting to recoup the loss; Pattern 4 captures a continuous feedback loop winning and then losing a bet; Pattern 5 captures a clawback behaviour where the bet was double the amount of previous bet following a lost bet.

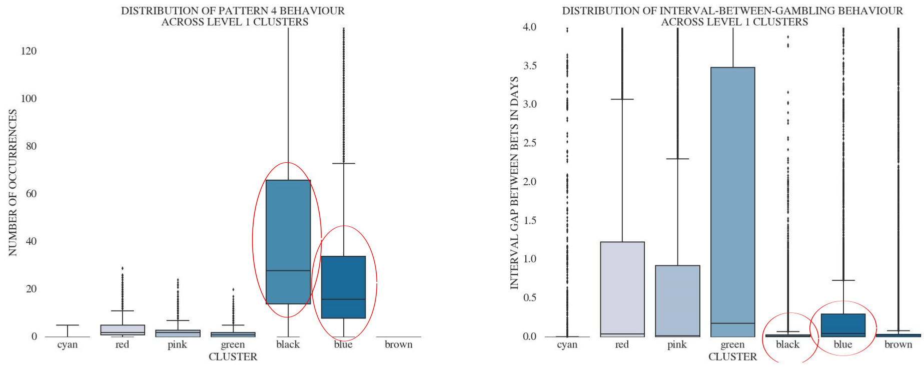

Fig. 2: A comparison of the distribution of Pattern 3 and Bet Interval. Similar distribution for other patterns is also observed (not shown above).

The time elapsed between bets (bet interval) was another marker of problem gamblers. The *event interval analysis* plug-in [13] was used with the ProM Tool to extract the bet interval for *each bet* placed by *every gambler* in *each cluster*. Additionally, the distribution of the median bet interval *per gambler* was extracted. Next, Kruskal-Wallis and Dunn's tests were applied to verify if the distributions of the bet intervals across the 7 clusters were different at *p*-value 0.05. The box-and-whisker graph (Fig. 2 - right) shows that the *blue* and *black* clusters had significantly lower bet interval values compared to other clusters. The exceptions being *cyan* and *brown* clusters, though they were discounted since their actual bet numbers were very low (e.g. mostly only bet once). We merged ('roll-up') the *blue* and *black* clusters as this was supported by the pair-wise Dunn's tests (which resulted in a *p*-value higher than 0.05, indicating insignificant differences).

These analyses allowed us to group and rank the severity of gambling patterns across the 7 clusters, resulting in 4 clusters that were statistically different (thus addressing RQ1). Our ranking exercise suggested that the *blue* and *black* clusters were more likely to contain problem gamblers, thus warranting further 'drilling-down' analysis into these two clusters. We performed a second-level clustering of these two clusters into another 7 clusters. For these second-level clusters, our focus was to establish statistical differences in terms of gambling clawback behaviour. We used the frequency distribution of those activities signifying loss-related clawback behaviours: `loss_<=1`, `loss_1to2`, and `loss_>=2`.

Similar to our earlier approach, we used the Kruskal-Wallis and Dunn's tests results on our second-level clustering to provide a more refined severity ranking of problem gamblers within the second-level clusters. Through this analysis, we addressed RQ2: we narrowed down the most likely problem gamblers to two clusters (from the second-level clusters) with a combined population size of 6,389.

Challenges Existing comparative analysis of processes [8, 9, 12] seemed to go no further than visual analysis of process models, or through some forms of multi-perspective visualisation techniques (e.g. [9]). However, such approaches lack theoretical vigour. As explained above, we addressed this issue by applying confirmatory statistical testing.Practically, we faced computational limitations during use of the event interval analysis [13] and BPMN Miner [1] plug-ins. We handled this by *further dividing* the event log of the largest cluster into three sub-logs. However, division of large XML-formatted event log was also a challenge as it contained more than 105 million XML elements/lines. We used basic, powerful, Linux-based text processing utility, such as `less`, and `sed` to cope with large XML files.

Evaluation. This case study relied heavily on clustering to identify various classes of gamblers.The main *challenge* therefore was on how to be sure that the clusters we obtained were indeed appropriate and meaningful. To do so, we had to assert that (1) the population within each cluster shared some unique characteristics and/or behaviours, and (2) the existence of those clusters was within reasonable expectation of the stakeholders.

For the former, we applied classification analysis on the combined first- and second-level clusters (12 clusters in total: 5 from the initial clusters and 7 clusters from the second-level clusters). We used features that *had not previously been used to generate the k-means clustering*, including the median bet interval values, the frequency of the 9 gambling patterns, the frequency of clawback behaviour activities, the total sum of money won, and the number of times a gambler won a bet. These features when used with the random forest algorithm (with 300 trees and 5-fold cross validation) generated a classification accuracy of 84.9% with a standard deviation of =/- 0.8%. These classification results, having used independent features from the clustering process, provided support that the clusters that were produced were relevant and meaningful.

The stakeholder also found the results to be reasonable. Some insights will require further study, e.g. the dominance of Pattern 3 and Pattern 4 (which suggested that gamblers had bet equal or less than their previous amount after they had won or lost the previous bet) which contradicted currently-understood risk-taking behaviour of problem gamblers where the expectation was for them to increase the amount of money bet following a loss. While further analysis may be required, there was an explanation for this: the dominance of Pattern 3 and Pattern 4 was evident when we took the all 12 clusters into account (which could mean that such behaviour is discriminatory for deciding if a gambler is a problem gambler or not). In the second level clustering, however, it was the frequency of high-risk clawback behaviour, i.e. the `loss_1to2` activity that is statistically different across all second-level clusters (thus in line with the expected behaviour).

4 Lessons Learned

During data processing stage, we found that it is *effective to combine unsupervised k-means clustering and aggregated case-level attributes* to slice a large data set was highlighted in this analysis. Doing so, we 'deflated' the data size from over 11 million events to just over 94 thousand lines of data (each line represents one gambler) that was further separated through with k-means clustering.

To estimate an optimal number of clusters that is conservative enough to reduce the error probability of not finding enough clusters and fine-grained enough to account for the complexity of the data, our case study shows that it can be achieved by *doubling the number of initially-suspected clusters and cross-referencing it with WSS analysis.*

We found *the non-parametric assumptions of Kruskal-Wallis tests, paired with Dunn's tests, made them flexible and useful* as, based on experience, data used for process mining analyses is *rarely* normally distributed in practice.

Our case study also shows that using *hierarchical clustering in combination with confirmatory statistics* is a suitable approach to apply the process cube concept [14]. Hierarchical clustering based on aggregated case-level attributes allowed us to 'slice and dice' event logs in an unsupervised manner, while confirmatory statistics informed us as to which groups of processes that should be 'rolled-up' (that is, those two groups whose p-value from the Dunn test indicates statistical similarities) and which to 'drill down' into. This is an alternative approach to the currently-prevailing approach to slicing processes that is based on control-flow perspective or simple attribute-based filtering [3, 8].

Finally, our case study demonstrates how one could apply classification analysis using features that were not used as input to the k-means clustering to provide some meanings to the clusters that we have managed to extract from the data. Recall that a detractor of k-means clustering is that it *will* produce as many clusters as parametrised. This is an important insight as previous work which used k-means clustering in a process mining case study [12] also applied classification analysis to provide meaning to each cluster. However, it was done using *the same features* as those used for clustering, which resulted in a very high accuracy rate but adds nothing more to our understanding of each cluster.

From a practical perspective, the Inductive Miner [5] implementationed in the ProM Tool is scalable to even the largest event log in our clusters. However, the process model was less so: it produced models that exhibited 'flower model' characteristics. By far, the Disco tool was still the most scalable tool in terms of extracting process models even on an event log of over 4GB in size.

We question the need for XML-formatted data in process mining. Raw data often come in a table-like format with fields whose meaning can be understood through discussions with stakeholders, thus lessening the need for the self-defining property of XML. XML files tend to be large and their processing is resource intensive. Our original CSV-formatted data (< 1.5 GB) was 'blown up' to over 5.8 GB after converting it to the XES format.

5 Related Work

Early process mining case studies [10,11,15] focused mainly on the application of standard process mining techniques, such as process discovery and performance analysis. Later process mining case studies [6–8] applied a combination of process and data mining techniques. Our case study fits closer to the latter: we applied a balanced amalgamation of process and data mining techniques. However, we ventured further by also using a multi faceted approach involving process mining and well-established confirmatory statistics.

This case study confirms some of the observations and experiences reported in other process mining studies. For example, the use of clustering techniques to split original event logs into smaller pieces was applied in our case study. However, instead of using trace/sequence clustering as in [2,10], we used k-means clustering [4] based on aggregated case-level features.

While an earlier case study [15] found that it was already feasible to conduct process mining analysis using the ProM Tool, we found that the quality and the robustness of the plug-ins in the ProM Tool to be highly variable with some being robust and scalable enough to handle large data sets (e.g. the Inductive Miner plug-in), while others tended to perform rather poorly (e.g. [1, 13]).

While our approach to process comparison was similar to other case studies, e.g. [8, 9], ours went further by studying the distribution of observed differences, and asserted their statistical significance using well-established confirmatory statistics. Most importantly, this paper reported new lessons learned that may be novel and helpful to other process mining practitioners.

6 Conclusion

We demonstrated how to apply a diverse and complementary set of techniques from the domains of process mining, data mining, and confirmatory statistics, as a unique expression of the process cube concept, to characterize and assert differences in gambling behaviours exhibited by more than 94 thousand gamblers in New Zealand. Most importantly, this case study reported a number of challenges and lessons learned that are novel and would likely be beneficial to other process mining practitioners.

References

1. R. Conforti, M. Dumas, L. Garca-Bauelos, and M. La Rosa. Beyond tasks and gateways: Discovering bpmn models with subprocesses, boundary events and activity markers. In *BPM*, volume 8659 of *LNCS*, pages 101–117. Springer, 2014.

2. J. De Weerdt, S vanden Brouckev, J. Vanthienen, and B. Baesens. Leveraging process discovery with trace clustering and text mining for intelligent analysis of incident management processes. In *IEEE CEC*, 2012.
3. C. C. Ekanayake, M. Dumas, L. Garcia-Banuelos, and M. La Rosa. Slice, mine and dice: Complexity-aware automated discovery of business process models. In *BPM*, volume 8094 of *LNCS*, pages 49–64, 2013.
4. J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 1979.
5. S.J.J. Leemans, D. Fahland, and W.M.P. van der Aalst. Discovering block-structured process models from incomplete event logs. In *Petri Nets*, volume 8489 of *LNCS*, pages 91–110. Springer, 2014.
6. J. Nakatumba. *Resource-aware business process management: analysis and support.* PhD thesis, Eindhoven University of Technology, 2013.
7. H Nguyen, M Dumas, M La Rosa, F.M Maggi, and S Suriadi. Mining business process deviance: A quest for accuracy. In *OTM 2014*, volume 8841 of *LNCS*, pages 436–445. Springer, 2014.
8. A. Partington, M.T. Wynn, S. Suriadi, C. Ouyang, and J. Karnon. Process mining for clinical processes: a comparative analysis of four australian hospitals. *ACM Trans. on Management Information Systems*, 5(4), 2015.
9. A. Pini, R. Brown, and M.T. Wynn. Process visualization techniques for multi-perspective process comparisons. In *AP-BPM*, volume 219 of *LNBIP*, 2015.
10. Á. Rebuge and D.R. Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Inf. Syst.*, 37(2):99–116, April 2012.
11. A. Rozinat, I.S.M de Jong, C.W. Gunther, and W.M.P. van der Aalst. Process mining applied to the test process of wafer scanners in ASML. *IEEE Trans. on System., Man, and Cybernetics, Part C*, 39(4):474 –479, 2009.
12. S Suriadi, R. S. Mans, M. T. Wynn, A Partington, and J Karnon. Measuring patient flow variations: A cross-organisational process mining approach. In *AP-BPM*, volume 181 of *LNBIP*, pages 43–58. Springer, 2014.
13. S. Suriadi, C. Ouyang, W.M.P. van der Aalst, and A.H.M. ter Hofstede. Event interval analysis: Why do processes take time? *Decision Support Systems*, 79:77–98, 2015.
14. W.M.P. van der Aalst. Process cubes: Slicing, dicing, rolling up and drilling down event data for process mining. In *Asia Pacific Conference on Business Process Management*, volume 159 of *LNBIP*, pages 1–22, 2013.
15. W.M.P. van der Aalst, H.A. Reijers, A.J. M. M. Weijters, B. F. van Dongen, A.K.A. Medeiros, M. Song, and H.M.W. Verbeek. Business process mining: An industrial application. *Information Systems*, 32:713–732, 2007.
16. M. L. van Eck, X. Lu, S.J.J. Leemans, and W.M.P. van der Aalst. PM^2: A process mining project methodologya. In *CAiSE*, volume 9097 of *LNCS*, pages 297–313. Springer International Publishing, 2015.
17. J. De Weerdt, S. vanden Broucke, J. Vanthienen, and B. Baesens. Active trace clustering for improved process discovery. *Knowledge and Data Engineering*, 25(12), 2013.