Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 1357 – 1366

Procedia
Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Prediction of Enterprise Purchases using Markov models in Procurement Analytics Applications

Adam Westerski[a]*, Rajaraman Kanagasabai [a], Jiayu Wong[b], and Henry Chang[b]

[a] *Institute for Infocomm Research, A*STAR, Singapore*
[b] *A*STAR Procurement Office, A*STAR, Singapore*

Abstract

Procurement is a set of activities and processes related to acquisition of goods and services through purchase orders placed by organization employees, from external contractors. This article describes practical experiments with procurement dataset of a major governmental organization in Singapore. In particular, we highlight the problems that emerge when trying to implement analytics for prediction of future purchases. The goal of such analytics is to deliver beneficial information to procurement office that plans and manages relationships with external sellers. In the article we describe the characteristics of the procurement dataset specifics and its implications on the future purchase problem that we attempt to solve using Markov chains model. Our analysis shows high diversity of purchase descriptions resulting in low ability to detect sequential patterns of purchasing officers. The solution presented in the article is additional dataset preprocessing involving use of hierarchical clustering. Our experiments with various similarity measures show an improvement allowing a practical deployment within our procurement analytics system prepared for the case study governmental organization.

Peer-review under responsibility of KES International

Keywords: prediction, procurement, analytics, purchase orders, clustering

* Corresponding author. Tel.: +65-64082681; fax: +65-67761378 .
E-mail address: adam-westerski@i2r.a-star.edu.sg

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.209

1. Introduction

Procurement is fast emerging as an area of economic importance. A recent report by Ardent Partners highlights that an average procurement department of an organization manages 60.6% of total enterprise spending [40]. Therefore, it is desirable that goods and services are procured at the optimal cost to the requester while adequately meeting the requirements of the organization. Unfortunately, this is not always the case as management of procurement processes in organizations is subject to a number of problems. Among those, the 2014 PwC Global economic crime survey reports a sharp rise in procurement fraud [41]. PwC identified that 29% of the companies are affected by at least one form of procurement fraud, making it the second most common economic crime. However, organizations loose resources not only due to fraud but also inefficient management of procurement. While analyzing spending of Italian public organizations, Bandiera et al. [39] propose distinction between active and passive waste. According to the findings, corruption related to active waste accounts only for 22%. The remaining loss of resources can be related to passive waste, which originates from lack of knowledge about the overall spending and their characteristics in an organization, lack of skills to reduce the costs or lack of incentives to do so, combined with excessive regulatory burdens.

To eliminate some of those aforementioned problems, traditionally, eyeball sampling has been used for compliance checking and fraud detection in procurement but this is a tedious and error-prone process. The increasing adoption of procurement software has given rise to the availability of transaction data and usage logs that lend themselves amenable for data analysis. Consequently, Business Intelligence software (e.g. Tableau [42]) has begun to be employed to assist in human analysis. In addition, tools that perform descriptive statistics have also been used for computing quantities such as mean and dispersion. Both these methods have been found to be useful for macro-level insights such as identifying departments or divisions of a company with maximum procurement spending, computing average expenditures and other statistical parameters such as quantiles or variances. They can also help in data visualization such as trend graphs, and data pre-processing tasks such as data validation. However, major procurement frauds such as bid rigging, collusion between suppliers, fraudulent payments via shell companies, etc. can be so subtle to make it difficult to detect by standard macro-level analysis [43].

Motivated by the issues mentioned above, we report on research made to enhance contemporary procurement software with predictive capabilities. We present results on future purchase order prediction method, which is a component in a larger framework developed as part of a governmental program on improving procurement management. The method has been tested and implemented as part of software called Procurement Tracker, which is used by one of the governmental agencies in Singapore to aid management of procurement and inspect purchases. The predictive purchases component enables procurement officers to see upfront expected orders that will be made by requesting officers, connect with the requesters and aggregate their orders before they are placed to save on purchase costs. Secondly, the predicted orders are an outcome of analyzed purchase history of requesters, therefore for requesters who show clear patterns in behavior, such predictions can be used to alert procurement officers of changes in the requesters' behavior, which can be the basis for further fraud investigation.

The structure of the article is as follows: firstly, we introduce the reader to work done so far in the area of procurement (see Sec. 2); later, we discuss the motivation for our work and the broader context of its application (see Sec. 3). Further, we proceed with the specifics of our approach by describing the input dataset, which is the foundation for our experiments (see Sec 4.1), necessary data pre-processing (see Sec. 4.2) and finally applied algorithmic approach (see Sec. 4.3). Next, we describe the setup of the evaluation experiments (see Sec. 5), and present their results and comment on lessons learned (see Sec 5.1). The article concludes with the overview of achievements and by pointing out directions for future work (see Sec. 6).

2. Related Work

As procurement has grown to be an important area of organization activity, there have been a number of works related to various problems of this domain: procurement fraud [1, 2, 3], aggregation of purchases [4], procurement issues related to underlying supply chain optimization [5, 6, 7]. Nevertheless these research initiatives are very selective and focused on specific scenarios, to our knowledge, not related to situation or problems of the organization described in this article.

The problems and solutions discussed by us in context of procurement can be also found in other domains. Fraud detection is a very popular research topic and according to a survey by Phua el al. [8] is most explored in context of financial fraud and credit card fraud (e.g. [9, 10, 11]). Similar problems have been also reported in many other areas such as telecommunications [12, 13] or insurance [14, 15]. Among those, the majority of proposed solutions revolve around employing data mining methods, which leverage statistical models [16] and probabilistic approaches [17] to create advanced machine learning algorithms [8]. Within the machine learning area, the vast majority of approaches are based on supervised learning algorithms, i.e. assume existence of well annotated set of past fraud cases. The popular techniques include use of neural networks [18, 19, 20] and Bayesian networks [21, 22, 23] or less frequently case-based reasoning [24, 25] or decision trees [26]. This assumption regarding the input dataset does not however apply to our scenario; therefore for our case there was a clear need to deliver an algorithm that would detect suspicious purchase orders without any prior knowledge on past fraud cases.

The particular path taken by us: next-order prediction, is not typically used to aid fraud detection, however there are a number of works that experiment with purchase prediction algorithms for other goals. Insurance purchase prediction [27, 28], website access prediction [29] and customer purchase prediction in commerce [30, 31] are some of the popular to name. Among those, similarly to us, Deshpande [29] experiments with different optimizations of Markov model to obtain better accuracy. In comparison to that work, we focus more on balance between accuracy and coverage as well as comparison of Markov to other approaches, while Deshpande proposes multiple Markov optimizations for increasing accuracy and compares them to each other in different application domains to our.

Similar predictive capability evaluations but expanded to particular application domains have been done for supply forecasting and demand aggregation [35]. This kind of research is closer to our second goal of predicting procurement actions for bulk purchases. However, apart of the features that we focus on, other parameters are being investigated. Jen et al. [32] propose prediction of purchase frequency and apply a different technique altogether to achieve this, i.e. Hierarchical Bayes. Among more closely related approaches, which apply Markov processes, typically the objective is to model item stock levels and predict demand per item in certain predefined time intervals [33, 34]. In comparison, we focus on predictions per requesting customer in an undefined time.

During the analysis of such prior work, we noted that many of those supply-demand models are extensive theoretical frameworks that assume existence of data with certain features and quality that is not always present in practical scenarios of organizations. In our study we were in a comfortable position of very tight collaboration with a governmental agency and full access to its data. Therefore, we were able to base the proposed approach on information that would be available in the system deployment scenario; and furthermore make decisions on application of algorithms after an extensive analysis of the quality of the available data. Insights from all of those phases are presented in the following sections of this article.

3. Approach

3.1. Dataset description

The presented research was done based on data produced as output of operations of a large governmental agency in Singapore over the course of four years, 2010- 2013 (inclusive). Throughout this period there were a total of 141,286 purchase orders (PO) recorded. Each of the purchase orders was placed by an employee of the agency (requesting officer) and later was subject of further approval by a procurement officer (approval officer). Therefore, each purchase order would have its creation date (day on which requesting officer submitted it to the system) and an approval date. Additionally, a constraint of the system was that a single purchase order would always be related to only one vendor who supplies the ordered goods or services of a given value expressed in local currency.

The purchase orders would further consist of purchase order items. Each of those can relate to different item or service and contain further details like: textual item description, quantity of items bought and unit price per single item. The dataset being subject of our analysis contained a total of 316,036 purchase order items. On average a single purchase order had attached 2.2 items, however 59% of purchase orders had only 1 order item and 97% having 10 or less. Within the remaining 3% the maximal recorded amount of purchase order items per purchase order was 164. This reflects the overall behavior of the agency employees and the policies in place that focused on simple orders, typically related to one type of good.

Beyond those general insights, we learnt that the key elements that influenced the performance and capabilities of prediction model were related to quantitative relationships between the four aspects of the procurement data (see Fig. 1): who (requester), what (item description), from who (vendor), and when (creation/approval date). Therefore, below we detail some of the key insights that later on will help to explain the results of obtained with our model.

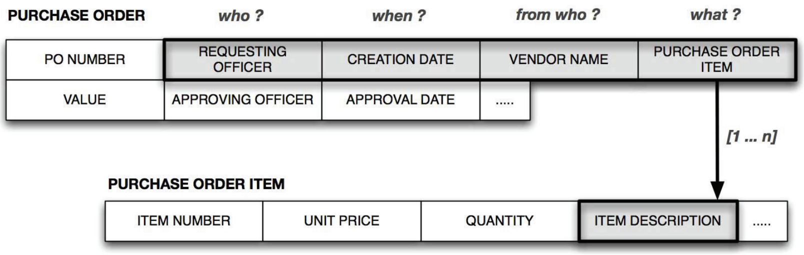

Fig. 1. Key elements of the single purchase order data vector.

Table 1 contains extended statistics of the dataset, which reveal that quite a significant amount of requesters participated in placing orders but at the same time for many cases the history of orders per requester is rather short, which could make it difficult or even impossible to perform any predictions.

Table 1. Breakdown of key statistics for the procurement dataset used in predictive model experiments.

Metric	Value	Metric	Value
# Purchase Order	141286	MIN/AVG/MAX #Order PER Requester	1/ 12.7/ 1002
# Purchase Order Item	316036	MIN/AVG/MAX #Vendor PER Requester	1/ 5.8/ 163
# Vendor	7887	MIN/AVG/MAX #Item PER Requester	1/ 11.7/ 807
# Requester	11312	MIN/AVG/MAX Creation Date Difference PER Requester (days)	0/ 25.92/ 491
# Approval Officer	594		
# Item	212652		

Further investigating this relationship between orders and requesters (see Fig. 2a), it can be observed that majority of requesters contributed to only about half of the dataset (requesters with less then 101 purchase order items), while the other part of the data relates to small portion of very active requesters that had some administrative roles and frequently placed orders. As it will be revealed in next sections, those were the truly interesting cases for us, since they had a rich history of orders that could be used for training of our predictive model.

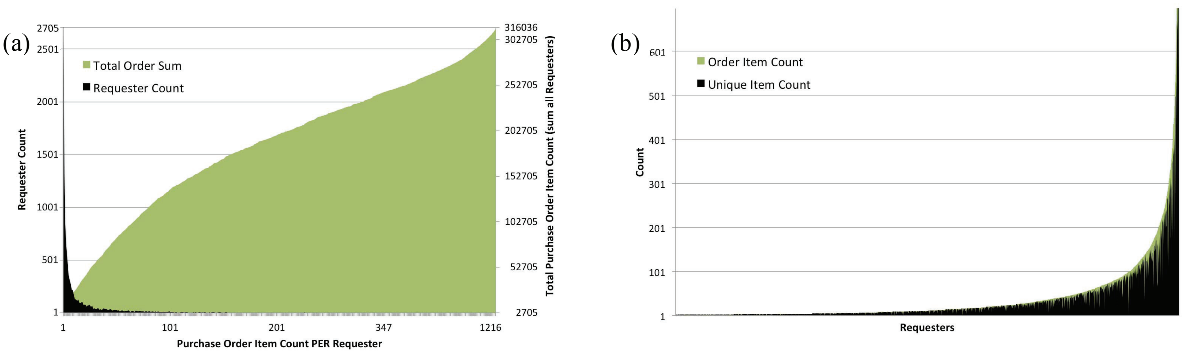

Fig. 2. (a) Distribution of purchase order items per requester: comparison of requester count (black) and sum of purchase order items (light green); (b) Comparison of unique item description count (black) and order item count (light green) per unique requester.

Aside of the requester analysis, we also investigated the diversity of orders placed to make some initial assumptions about the difficulty to perform predictions with relation to textual descriptions of purchased items. Figure 2b shows that regardless of volume of their activity, requesters seem to repeat their purchases very little, as in most cases the amount of unique descriptions is very similar to total orders made. Based on this observation, our first hypothesis (H1) was that any sort of prediction for a meaningful amount of requesters could be very difficult due to lack of any observable purchasing patterns. However, in a preliminarily experiment, we manually analyzed textual item descriptions for orders made by *9 requesting officers* picked on random (total of *1000 purchase order items*). It turned out that often the same or similar items were described only with a slightly different text. This was a result of requesting officers having to input the item descriptions manually without any shared dictionary or index. Based on this observation, we formed another hypothesis (H2) that the base dataset could be pre-processed using text-clustering techniques to match similar descriptions and greatly improve the final predictive capabilities. The details of the algorithms used for both data pre-processing and actual predictions are discussed in the next sections.

3.2. Data preprocessing – clustering of purchase orders

During the data pre-processing phase we applied *hierarchical clustering* algorithm and calculated similarity between textual descriptions using q-gram distance [36], i.e. sum of absolute differences between q-gram vectors (substrings of length q) of compared strings [37]. Furthermore, we analyzed a number of different similarity thresholds relating to biggest distance allowed for textual descriptions in the same cluster. Apart of such clustering, we experimented with additional data pre-processing steps: (a) clustering text as it was input originally by requesters or by transforming all upper case characters into lower case; (b) setting a string length threshold below which textual descriptions would be excluded from clustering. The (a) scenario was related to observation that some of the item descriptions were input fully with upper case characters, while others with lower case. In the second scenario (b), our attention to limiting string length was based on fairly big diversity of this parameter across the dataset. We determined that shortest item description length for an individual requester had a medium positive correlation with clustering performance parameters (precision, recall, f-measure); and the longest item description length for requester was related to dendrogram height (due to string similarity algorithm choice) and therefore had an impact on clustering performance when using with different similarity thresholds. Our concern was that large q-gram similarity thresholds would produce biased clustering results for requesters with large number of short item descriptions, and that it would have a significant impact on overall final results for the entire dataset.

3.3. Purchase prediction algorithms

The dataset analysis has shown a significant relationship between requester and parameters that could influence purchase prediction. Therefore, all our approaches focused on predicting item descriptions for a single requester at a time. We evaluated a number of algorithms starting with simple random sampling, through probability distribution analysis, ending with sequence analysis using Markov chains. Given the last purchase of a requester, we applied each of those algorithms to calculate: *single next purchase prediction* and *prediction of multiple purchases*. The goal was to provide prediction for the biggest amount of requesters thus covering the biggest portion of future purchases. Given the applied nature of our work, the desire was to embed configurability into algorithms so that the end user could consciously maximize on performance for one of those aforementioned parameters at the cost of others.

Random Sampling: the simplest solution, treated as baseline reference for others. In the training phase this algorithm collects all unique item descriptions from the requester order history. During the test phase a single value is picked from this set using random number generator. Therefore, the last purchase made by requester does not have any effect on prediction for this algorithm and both *next purchase prediction* and *multiple purchase prediction* are handled in similar manner by repeating the described random picking process for multiple predictions.

Probability Distribution: in the training phase for this algorithm we first calculate the Probability Density Function (PDF) for the predicted feature limiting to orders of a given requester. Next, based on the PDF, we create a Cumulative Distribution Function (CDF). This function would serve as a reference for prediction of future orders. Moving forward, during the test phase, we calculate a random variable value between 0 and 1. This value would be passed as argument to the CDF function and the corresponding feature value would be extracted as final prediction.

Simple Sequential Sampling: used as an additional reference specifically for prediction of multiple purchases. Like Random Sampling during the training phase it collects all unique description values for a given requester. However, this time, those values would be recorded in the same sequence as they occurred in the order history. During the test phase, depending on the predicted series length, the prediction would be sequence of orders starting from the first recorded value from training phase. If the predicted series length exceeded amount of unique feature values of a requester, then the prediction would continue from the start of the trained sequence all over again.

Markov Chain: contrary to previous approaches, this solution provides a prediction dependent on previous state recorded for the requester. We focused on evaluating first order Markov chains, i.e. decision about next state was made based on a single previous state. We did not test k-order chains due to nature of the dataset highlighted earlier, i.e. large amount of requesters with little orders. Our anticipation was that such data structure would lead to situation where increasing order of Markov chains would produce satisfactory results for few requesters. Such conclusions have been reached based on experiments of Deshpande [27] who shows that k-order chains decrease drastically the dataset coverage and algorithm runtime performance - both being key parameters for our implementation.

During the training phase for this algorithm, we build a transition matrix that models probability of moving between states of the Markov chain. In our case, the states are unique item descriptions taken from order history of a requester. Transition between those states occurs when a purchase with a particular feature value is followed by another different feature value. If purchases with the same feature values happen one after another the state does not change. During the construction of this matrix, the purchases of a requester would be ordered according to their creation date, starting from the earliest and ending with the latest. The probability for transition from a given state to another is calculated based on amount of times a given transition was observed in training data in relation to total amount of time any transition from the origin state has occurred. Therefore, each state in the matrix has it's own probability distribution similar as in earlier approach using CDF but this time dependent on previous state. Given this transition matrix, during the test phase the feature value of the last performed order would be used as input to identify the starting state in the matrix and the probability distribution to be used. The predicted feature value for next purchase would be determined in a similar procedure as in the Probability Distribution approach.

On top of this base solution involving Markov chains we also added two optional optimizations, both applied at the training phase stage: *creation date optimization*; and *frequency count optimization*. The first, *creation date optimization*, involves the granularity of creation date for purchase order. If this optimization is used orders made on the same day are not treated a consecutive but concurrent. Therefore, the order made on previous day has a transition to all orders made on next day with the same probability for each. The second Markov chain optimization, *frequency count optimization*, relates to reducing the size of Markov chain for a requester. This operation is done through removing all transitions with frequency count below a certain selected threshold. Furthermore, if all transitions of a requester do not meet the threshold the requester is ignored and no predictions are made. This optimization is aimed to give the capability to manipulate the minimum activity of requester required to provide predictions.

Given the presented approach and proposed optimizations, our hypothesis was that application of Markov chains would improve over all of solutions with respect to most active requesters having small average time distance between their orders. In the next section, we present the evaluation setup and experiments verifying this hypothesis.

4. Evaluation: clustering and data pre-processing

In the first stage of evaluation we adjusted the parameters for data pre-processing to obtain the best input for the predictive algorithms used later. We experimented with hierarchical clustering by evaluating six different similarity thresholds for the q-gram algorithm: 8, 16, 24, 32, 40 and 48. On top of that we tested bi-gram and tri-gram analysis for each of those thresholds. In each of the aforementioned scenarios we used input data being: (a) original textual item description input; (b) item description converted into lower case. Secondly, we evaluated if overall clustering performance would change if we excluded descriptions of character count below 20. In all of those setups the evaluation was done based on comparison of a manually annotated dataset of 2000 order items relating to 44 requesting officers. The officers for evaluation were picked on random from requester groups with different dendrogram heights representative for the entire spectrum of the dataset (for the test dataset the min. dendrogram height was 10, max. height 77, average height 53.3, and standard deviation 20.4). During the evaluation we calculated three parameters: precision, recall and f-measure. Due to the final goal of applying in prediction algorithms, precision would be equally important to recall, therefore the key parameter we paid most attention was optimization of f-measure. Those performance parameters were calculated based on pair-wise comparison of cluster assignments done by our algorithm in comparison to the test set (as described by Manning et al. [38]):

$$TP = \ count\ of\ assignements\ of\ two\ similar\ item\ descriptions\ to\ the\ same\ cluster$$

$$FP = \ count\ of\ assignements\ of\ two\ dissimilar\ item\ descriptions\ to\ the\ same\ cluster$$

$$FN = \ count\ of\ assignements\ of\ two\ similar\ item\ descriptions\ to\ different\ clusters$$

$$precision = \frac{TP}{TP+FP} \qquad recall = \frac{TP}{TP+FN} \qquad f-measure = 2 \cdot \frac{precision \times recall}{precision + recall}$$

We used the above metrics to properly tune the clustering algorithm and identify the best settings for clustering item descriptions. Figure 3 shows comparison of results for different clustering approaches in relationship to similarity threshold value.

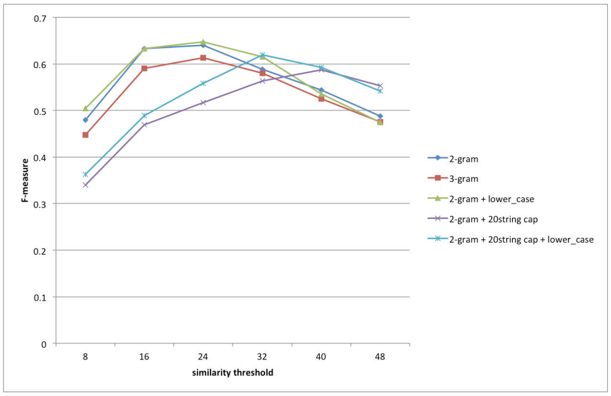

Fig. 3. Evaluation results of clustering in relation to similarity threshold.

As it can be observed, the best results were obtained for bi-gram similarity measure with lower case item descriptions transformation and q-gram 24-similarity threshold; we used this method in conjunction with our Markov chain algorithm to further evaluate item description predictions, as described in next section.

5. Evaluation: future purchase prediction

In the second part of our experimentation we tested the results of purchase prediction algorithm with the best clustering setup and without any clustering for comparison. Prediction tests were done separately for every requester and afterwards the results were averaged across requesters to give the final result. Half of the requester order data was used as training set, and half as test set. For every requester we assumed a minimum requirement of 1 order in the training set, and at least 2 orders in the test set (aside of test prediction result, 1 order in the test set would always be necessary to determine the current state and serve as prediction reference in Markov approach). Requesters not meeting this requirement would be completely ignored and not taken into account during the experiment.

Using the previously described algorithms we run the evaluation in two test scenarios: (a) *next-order prediction*: for a particular requester, given a certain last performed order information, predict the selected feature value in the next order; (b) *series of orders prediction*: for a particular requester, given a certain last performed order information, predict the selected feature value for a given amount of next orders coming one after another in an undefined time period. For evaluating the quality of the *next-order prediction* approach we used accuracy parameter, which for a single test would take value 1 if predicted feature value matched the expected value and 0 if not. The prediction accuracy for a given requester would be expressed as average of all predictions tested for this requester.

During the *series of orders prediction* experiment we additionally split the test set into series of predefined length and for each used precision and recall parameters defined as follows:

$$precision = \frac{count\ of\ correctly\ predicted\ feature\ values}{count\ of\ all\ uniqe\ predicted\ feature\ values} \qquad recall = \frac{count\ of\ correctly\ predicted\ feature\ values}{count\ of\ all\ unique\ feature\ values\ in\ test\ subset}$$

In the above equations, the correctly predicted feature values were those, which would be mentioned in the predicted series of orders as well as exist in the series of orders extracted from the test subset (regardless of order within the series). Contrary to clustering experiments, in our evaluation of future purchase prediction, we aimed to optimize the solution to give best possible precision, as in the end solution the requirement for software was that we could afford to miss some of the purchases but should not provide faulty predictions. For all combinations of test scenarios and algorithms we analyzed the results in different ranges of algorithm performance (i.e. ranges of accuracy or precision/recall depending on approach). This was related to earlier described requirements for Procurement Tracker where the end user would be able to adjust the required prediction performance.

Taking into account such evaluation setup, we started with experiments for unprocessed input to verify if our concerns were true to the actual state and to have a comparison as to how much we could improve by using additional data pre-processing methods. We tested the *next-order prediction* using the random sampling, CDF and Markov; followed by *20-order series prediction* experiment with all four algorithms. For *next-order prediction* Markov approach gave best accuracy results out of all algorithms, however overall the results were unsatisfying for a practical application. For requesters where prediction was possible (accuracy > 0) Markov approach got 0.31 accuracy with 6.64% dataset coverage without clustering; and 0.26 accuracy with 34.27% coverage for clustering enabled. Although, inclusion of clustering lowered the accuracy, it significantly helped to provide predictions for more requesters.

In case of *20-order series prediction* it could be clearly noted that loosening the requirements for prediction timeframe provided a significant improvement in terms of performance (see Table 2). However, for the unclustered input the coverage of the dataset remained still far below our expectations. Therefore, in the final approach, we compared those results with input dataset optimized using previously discussed techniques.

Table 2. Summary of key experiments for series of orders item description prediction.

Setup	Ignored Requesters (% of all requesters)	AVG Precision/Recall for Requester (Requester Count / % of Dataset orders / precision / recall)		
		Precision >= 0	Precision > 0	Precision > 0.5
0.5 train+ Markov+ 20 order set	98.16%	212 / 15.04% / 0.34 / 0.09	96 / 5.70% / 0.74 / 0.19	67 / 1.61% / 0.97 / 0.26
0.5 train+ CDF+ 20 order set	35.58%	6438 / 97.50% / 0.04 / 0.03	1319 / 49.41% / 0.20 / 0.17	120 / 0.88% / 0.95 / 0.80
0.5 train+ Random Sampling+ 20 order set	35.58%	6438 / 97.50% / 0.04 / 0.03	1258 / 46.28% / 0.19 / 0.17	110 / 0.58% / 0.94 / 0.84
0.5 train+ Sequence Prediction+ 20 order set	35.58%	6438 / 97.50% / 0.04 / 0.04	1367 / 40.10% / 0.17 / 0.19	98 / 0.54% / 0.93 / 0.91
0.5 train+ Markov+ 20 order set+ clustering	**72.06%**	**2356 / 78.16% / 0.32 / 0.08**	**1235 / 52.47% / 0.61 / 0.15**	**600 / 12.88% / 0.96 / 0.24**
0.5 train+ CDF+ 20 order set + clustering	37.39%	6214 / 97.05% / 0.15 / 0.13	3134 / 83.76% / 0.29 / 0.25	382 / 2.59% / 0.91 / 0.75
0.5 train+ Random Sampling+ 20 order set + clustering	37.39%	6214 / 97.05% / 0.00 / 0.00	14 / 1.05% / 0.01 / 0.02	0 / 0.00% / 0.00 / 0.00
0.5 train+ Sequence Prediction+ 20 order set + clustering	37.39%	6214 / 97.05% / 0.00 / 0.00	6 / 0.31% / 0.02 / 0.03	0 / 0.00% / 0.00 / 0.00

Inspecting Table 2 it can be seen that Markov experiment with the dataset optimization greatly improves with regard to coverage issue at a small cost of slightly decreased precision. In comparison to other algorithms tested

(CDF, Random Sampling and Sequence Sampling) even the Markov experiment with unclustered input greatly excels in terms of precision yet falling behind in terms of coverage, an issue that is partially addressed by clustering.

6. Conclusions

Analyzing the experiments and results presented, it can be noted that in applied analytics solutions the choice of the final predictive algorithm is but one of many tasks. According to our experiences coming from the described work with industrial partners on procurement analytics, the key difficulties prove to be: proper input dataset analysis to understand the horizon of practical deployment possibilities; and dataset pre-processing that may improve greatly on whatever can be achieved with raw data. As it can be seen in evaluation results, without taking those issues into account, all tested solutions gave unsatisfactory results.

In the particular context of our experiments, we have shown that clustering of item descriptions greatly improved on the predictive capabilities of the algorithms used. However, one has to keep in mind that this particular combination of clustering and predictive algorithms requires proper balancing between how generic the predictions get (increase on clustering threshold) and how good the final accuracy of the predictive algorithms is. In our case, discovering the exact guidelines for this and further improving on performance of both algorithms remains the topic of future work.

References

1. Ruping S, Punko N, Gunter B, Grosskreutz H. Procurement Fraud Discovery using Similarity Measure Learning. *Tran. CBR* 2008; **1**(1): 37-46.
2. Calafato A, Colombo C, Pace GJ. A Domain Specific Property Language For Fraud Detection To Support Agile Specification Development. *Computer Science Annual Workshop (CSAW 2014)*. Malta; 2014.
3. Jans M, Lybaert N, Vanhoof K. Internal fraud risk reduction results of a data mining case study. *International Journal of Accounting Information Systems* 2010; **11**(1): 17–41.
4. Wang G, Miller S. Intelligent Aggregation of Purchase Orders in e-Procurement. *Proceedings of the 2005 Ninth IEEE International EDOC Enterprise Computing Conference (EDOC'05)*. Australia; 2005.
5. Buffett S, Scott N. An Algorithm for Procurement in Supply-Chain Management. *AAMAS 2004 Workshop on Trading Agent Design and Analysis*. New York, NY; 2004.
6. Yuan H, Sun X, Jing Y, Pan W, Ren T. Design of Material Ordering Strategy in Procurement Supply Chain. *Chinese Control and Decision Conference (CCDC)*. Mianyang, China; 2011.
7. Mori M, Kobayashi R, Samejima M, Komoda N. An Evaluation Method of Reduced Procurement Risks by Decentralized Ordering in Supply Chain. *11th IEEE International Conference on Industrial Informatics (INDIN)*. Bochum, Germany; 2013.
8. Phua C, Lee V, Smith K, Gayler R. *A Comprehensive Survey of Data Mining-based Fraud Detection Research*. Clayton School of Information Technology, Monash University; 2005
9. Kauderer H, Nakhaeizadeh G. The Effect of Alternative Scaling Approacheson the Performanceof Different Supervised Learning Algorithms. *An Empirical Study in the Case of Credit Scoring. AAAI Technical Report WS-97-07*. 1997.
10. Haimowitz IJ, Schwarz H. Clustering and Prediction for Credit Line Optimization. *AAAI Technical Report WS-97-07*. 1997.
11. Hanagandi V, Dhar A, Buescher K. Density-Based Clustering and Radial Basis Function Modeling to Generate Credit Card Fraud Scores. Proceedings of the IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering. New York City, NY; 1996.
12. Cahill M. Chen F, Lambert D, Pinheiro J, Sun D. Detecting Fraud in the Real World. *Handbook of Massive Datasets*. Norwell, MA, USA: Kluwer Academic Publishers; 2002. p. 911-930
13. Cortes C, Pregibon D, Volinsky C. Computational Methods for Dynamic Graphs. J*ournal of Computational and Graphical Statistics* 2003; **12**(4): 950-970.
14. Bentley P. Evolutionary, my dear Watson: Investigating Committee-based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2000)*. 2000
15. Cox E. A Fuzzy System for Detecting Anomalous Behaviors in Healthcare Provider Claims. In: Goonatilake S, Treleaven, P, editors. *Intelligent Systems for Finance and Business*. John Wiley and Sons Ltd; 1995, p. 111-134
16. Bolton RJ, Hand DJ. Statistical Fraud Detection: A Review. *Statistical Science* 2002; **17**(3): 235-255.
17. Hollmen J. *Probabilistic Approaches to Fraud Detection*. Helsinki University of Technology; 1999

18. Ghosh S, Reilly D. Credit Card Fraud Detection with a Neural Network. *Proceedings of 27th Hawaii International Conference on Systems Science*. 1994. p. 621-630.
19. Barse E, Kvarnstrom H, Jonsson E. Synthesizing Test Data for Fraud Detection Systems. *Proceedings of the 19th Annual Computer Security Applications Conference*. 2003. p. 384-395.
20. Syeda M, Zhang Y, Pan Y. Parallel Granular Neural Networks for Fast Credit Card Fraud Detection. *Proceedings of the 2002 IEEE International Conference on Fuzzy Systems*. 2002
21. Ezawa K, Norton S. Constructing Bayesian Networks to Predict Uncollectible Telecommunications Accounts. *IEEE Expert October* 1996; **11**(5): 45-51.
22. Maes S, Tuyls K, Vanschoenwinkel B, Manderick B. Credit Card Fraud Detection using Bayesian and Neural Networks. *Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies*. 2002.
23. Viaene S, Derrig R, Baesens B, Dedene G. A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. *Journal of Risk and Insurance* 2002; **69**(3): 373-421.
24. Wheeler R, Aitken S. Multiple Algorithms for Fraud Detection. *Knowledge-Based Systems* 2000; **13**(3): 93-99.
25. Curet O, Jackson M. Tackling cognitive biases in the detection of top management fraud with the use of case-based reasoning. *Proceedings of Expert Systems '95*. 1995. p. 223-236.
26. Fan W. Systematic Data Selection to Mine Concept- Drifting Data Streams. *Proceedings of SIGKDD04*. 2004. p. 128-137.
27. Rahimie Bin Md Noor M, mat Isa Z. Predicting Number of Purchasing Life Insurance Using Markov Chain Method. *Applied Mathematical Sciences* 2014; **82**(8) 2014
28. Bozzetto J-F, Tang L, Thomas LC, Thomas S. *Modelling the Purchase Dynamics of Insurance Customers Using Markov Chains*. School of Management, University of Southhampton; 2005
29. Deshpande M, Karypis G. Selective Markov Models for Predicting Web-Page Accesses. *ACM Transactions on Internet Technology* 2004; **4**(2): 163-184.
30. Bertsimas DJ, Mersereau AJ, Patel NR. Dynamic Classification of Online Customers. *Proceedings of the Third SIAM International Conference on Data Mining*. San Francisco, CA, USA; 2003.
31. Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing Personalized Markov Chains for Next-Basket Recommendation. *Proceedings of the 19th international conference on World Wide Web (WWW 2010)*. New York, NY, USA; 2010.
32. Jen L, Chou C, Allenby GM. A Bayesian Approach to Modeling Purchase Frequency. *Marketing Letters* 2003; **14**(1): 5-20.
33. Galliego G, Hu H. Optimal Policies for Production/Inventory Systems with Finite Capacity and Markov-Modulated Demand and Supply Processes. *Annals of Operations Research* 2004; **126**: 21–41.
34. Berling P, Xie, Z. Approximation algorithms for optimal purchase/inventory policy when purchase price and demand are stochastic. *OR Spectrum* 2014; **36**: 1077–1095.
35. Murty KG. Forecasting for Supply Chain and Portfolio Management. In: Neogy SK, Bapat RB, Das AK, Parthasarathy T, editors. *Mathematical Programming and Game Theory for Decision Making*. 2008.
36. Ukkonen E. Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science* 1992; **92**(1): 191–211.
37. Van der Loo MPJ. The stringdist Package for Approximate String Matching. *R Journal* 2014; 6(1): 111-122.
38. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge University Press; 2008.
39. Bandiera O, Prat A, Valletti T. Active and Passive Waste in Government Spending: Evidence from a Policy Experiment. *The American Economic Review* 2009; 99(4): 1278-1308
40. Bartolini A. *CPO 2011: Innovative Ideas for the Decade*. Ardent Partners Research; 2011
41. PwC Global Economic Crime Survey 2014. 2014.
42. Tableau Software, Available at: http://www.tableau.com/
43. Guile R. Procurement Fraud. In: Doig A. *Fraud: The Counter Fraud Practitioner's Handbook*. Gower Publishing; 2013.