Mining High Impact Exceptional Behavior Patterns*

Longbing Cao[1], Yanchang Zhao[1], Fernando Figueiredo[2], Yuming Ou[1], and Dan Luo[1]

[1] Faculty of Information Technology, University of Technology, Sydney, Australia
[2] Centrelink, Australia
{lbcao,yczhao,yuming,dluo}@it.uts.edu.au,
fernando.figueiredo@centrelink.gov.au

Abstract. In the real world, exceptional behavior can be seen in many situations such as security-oriented fields. Such behavior is rare and dispersed, while some of them may be associated with significant impact on the society. A typical example is the event September 11. The key feature of the above rare but significant behavior is its high potential to be linked with some significant impact. Identifying such particular behavior before generating impact on the world is very important. In this paper, we develop several types of high impact exceptional behavior patterns. The patterns include frequent behavior patterns which are associated with either positive or negative impact, and frequent behavior patterns that lead to both positive and negative impact. Our experiments in mining debt-associated customer behavior in social-security areas show the above approaches are useful in identifying exceptional behavior to deeply understand customer behavior and streamline business process.

1 Introduction

High impact exceptional behavior refers to customers' behavior, for instance, actions taken by them, aiming or leading to specific impact on certain business or societies. The *impact* can take form of an event, disaster, government-customer debt or other interesting entities. For instance, in social security, a large volume of isolated fraudulent and criminal customer activities can result in a large amount of government customer debt. Similar problems may be widely seen from other emerging areas such as distributed criminal activities, well-organized separated activities or events threatening national and homeland security, and self-organized computer network crime [5]. Activities or events in traditional fields such as taxation, insurance services, telecommunication network, drug-disease associations, customer contact center and health care services may also result in impact on related organization or business objectives [8]. Therefore, it is important to specifically discover such impact-oriented behavior to find knowledge about what types of behavior is exceptionally associated with target impact of high interest to management.

* This work is sponsored by Australian Research Council Discovery Grant (DP0773412, DP0667060) and Linkage Grant (LP0775041).

T. Washio et al. (Eds.): PAKDD 2007 Workshops, LNAI 4819, pp. 56–63, 2007.

There are the following characteristics of impact-targeted exceptional behavior. First, impact-targeted exceptional behavior specifically refers to those behavior itself, rather than behavior outcomes such as events, which has resulted or will result in big impact on the running of a business. Second, impact-targeted exceptional behavior is normally rare and dispersed in large customer populations and their behavior. They present unbalanced class and itemset distributions.

In this paper, we present lessons learnt in discovering low frequent and sequential exceptional behavior but associated with high impact in the social security domain. First, a strategy involving domain knowledge is discussed to partition and re-organize unbalanced data into *target set*, *non-target set* and *balanced set*, and construct impact-targeted activity baskets or sequences individually. We then mine exceptional behavior frequently leading to either *positive* (say {*P*-->T}, T refers to impact) or *negative* [6] (e.g., {*P*--> $\overline{\mathrm{T}}$ }) impact in unbalanced data. *Impact-contrasted* exceptional behavior *patterns* identify significant difference existing in two frequent patterns discovered on the same behavior basket or sequence in target set and non-target set, respectively.

We illustrate our approaches through analyzing exceptional behavior patterns leading to debt and non-debt in debt-related social-security activity data in Centrelink [1]. The outcomes of this research are of interest to Centrelink for understanding, monitoring and optimizing government-customer contacts, to prevent fraudulent activities leading to debt, and to optimize social security processes, therefore improving government payment security and policy objectives.

2 Preparing Exceptional Behavior Data

In practice, high impact exceptional behavior is a very small fraction of the whole relevant behavior records. It presents *unbalanced* [7] *class distribution* and *unbalanced itemset distribution*. Such unbalanced data makes it difficult to find useful behavior patterns due to many reasons.

To deal with the imbalance of exceptional behavior classes and itemsets, as shown in Table 1, unbalanced exceptional behavior data is organized into four data sets: original *unbalanced set*, *balanced set*, *target set*, and *non-target set*. For instance, all exceptional behavior instances related to debt in social security area are extracted into debt activity set, while those unlikely linked to debt go to non-debt set. A balanced activity set is to extract the same number of non-debt activity baskets/sequences as that of debt-related ones. The partition and re-organization of unbalanced activity data can deduce the imbalance effect, boost impact-oriented exceptional behavior, and distinguish target and non-target associated instances. In this way, impact-targeted exceptional behavior patterns easily stand out of overwhelming non-impact itemsets.

In social security government-customer contacts, the method to build activity basket/sequence is as follows. For each debt, those activities within a time window immediately before the occurrence of a debt are put in a basket/sequence. The time window is then moved forwards targeting the second debt for building another basket/sequence starting at the new occurrence of the debt.

Table 1. Partitioning unbalanced exceptional behavior data into separated sets

Set	Description
Unbalanced set	The original data set including both target and non-target exceptional behavior with unbalanced class distribution
Balanced set	A boosted data set including both target and non-target exceptional behavior with balanced class distribution
Target set	A data set solely including data of target-oriented class
Non-target set	A data set solely including data of non-target-oriented class

Furthermore, it is a strategic issue to determine the size of sliding time window. Domain knowledge, descriptive statistics and domain experts [2] are used to determine the window size. There may be varying methods to build such sliding window separating behavior instances in terms of the occurrence of an impact (1) with *changing window size*, namely the window covers all behavior instances between two impacts, (2) with *fixed window size*, namely the left hand side of the window always covers the impact, (3) with *tilt window*, namely the size can be either fixed or changing, the left hand side of the window always stop at a target impact, while the window may cover a long time period with coarser granularity for earlier behavior and finest for the latest.

For instance, Figure 1 shows two strategies, where a_i (i=1,…, m) denotes a normal behavior and d_j (j=1,…, n) is a debt closely associated with a series of behavior instances. In Changing Window mode, all behavior instances between the occurrences of two debts are packed into one window. This mode is more suitable for those applications with a frequent targeted impact. For instance, debt d_3 window includes behavior instance sequence $\{a_{14}, a_{15}, a_{16}, a_{17}, a_{18}, d_3\}$. Fixed Window mode fixes the length of the sliding time window, and packs all behavior instances in the window exactly before the occurrence of an impact into one window, say $\{a_8, a_9, a_{10}, a_{11}, a_{12}, a_{13}, d_2\}$. In Tilt Window mode, behavior instances happened in early time are considered but with low weight. This can be through sampling. For instance, for the scenario in Figure, we build sequence $\{a_6, a_9, a_{11}, a_{13}, a_{14}, a_{15}, a_{16}, a_{17}, a_{18}, d_3\}$.

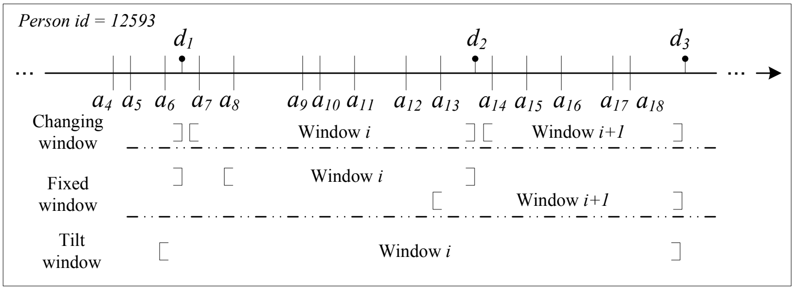

Fig. 1. Modes for constructing exceptional behavior window

On the other hand, *negative* impact targeted baskets/sequences are useful for contrast analysis. The strategy we use for building non-debt related customer behavior in social security area is to match with the positive impact scenario. Using the Fixed

Window mode, we can insert a non-debt impact into the behavior sequence if there is no debt happened then.

3 Mining High Impact Exceptional Behavior Patterns

3.1 Positive/Negative Impact-Oriented Exceptional Behavior Patterns

An impact-oriented exceptional behavior pattern is in the form of {*P*-->*T*}, where the left hand *P* is a set or sequence of exceptional behavior and the right hand of the rule is always the target impact *T*. Based on the impact type, both *positive* and *negative* impact oriented exceptional behavior patterns may be discovered.

Definition 1. *Frequent positive impact-oriented exceptional behavior patterns* (*P* --> *T*, or $\overline{P}$ --> *T*) refer to those exceptional behavior more likely leading to positive impact, resulting from either the appearance (*P*) or disappearance ($\overline{P}$) of a pattern.

Definition 2. *Frequent negative impact-oriented exceptional behavior patterns* (*P* --> $\overline{T}$, or $\overline{P}$ --> $\overline{T}$) indicate the occurrence of negative impact ($\overline{T}$), no matter whether an activity itemset happens or not.

In unbalanced set, a frequent impact-oriented exceptional behavior sequence leads to positive impact *T*: {*P* --> *T*} if *P* satisfies the following conditions:

- *P* is frequent in the whole set,
- *P* is far more frequent in target data set than in non-target set, and
- *P* is far more frequent in target set than in the whole data set. To this end, we define the following interestingness measures.

Given an activity data set *A*, based on exceptional behavior sequence construction methods, a subset *D* of *A* consists of all exceptional behavior baskets/sequences which are associated with positive impact, while the subset $\overline{D}$ includes all exceptional behavior baskets/sequences related to negative impact. For instance, in social security network, an exceptional behavior itemset *P* ($P = \{a_i, a_{i+1}, \ldots\}$, $a_i \in A$, $i=0, 1, \ldots$) is associated with debt *T*: {*P*--> *T*}. The count of debts (namely the count of sequences enclosing *P*) resulting from *P* in *D* is |*P*, *D*|, the number of debts resulting from *P* in *A* is |*P*, *A*|, |*P*, $\overline{D}$| is the count of non-debts resulting from *P* in non-debt subset $\overline{D}$, |*A*| is the count of debts in set *A*, and |$\overline{D}$| is the count of non-debts in set $\overline{D}$. We define the following interestingness measures.

Definition 3. The *global support* of a pattern {*P*--> *T*} in activity set *A* is defined as $Supp_A(P,T)$ = |*P*, *A*|/|*A*|.

$Supp_A(P,T)$ reflects the global statistical significance of the rule {*P*--> *T*} in unbalanced set *A*. If $Supp_A(P,T)$ is larger than a given threshold, then *P* is a frequent exceptional behavior sequence in *A* leading to debt.

Definition 4. The *local support* of a rule {*P*--> *T*} in target set *D* is defined as $Supp_D(P,T)$ = |*P*, *D*|/|*D*|. On the other hand, the local support of rule {*P*--> $\overline{T}$ } in exceptional behavior set $\overline{D}$ (i.e., non-debt exceptional behavior set) is defined as $Supp_{\overline{D}}(P,\overline{T})$ = |*P*, $\overline{D}$|/|$\overline{D}$|.

Definition 5. The *class difference rate* $Cdr(P,|^{D}_{\bar{D}})$ of P in two independent classes D and $\bar{D}$ is defined as

$$Cdr(P,|^{D}_{\bar{D}}) = Supp_D(P,T) / Supp_{\bar{D}}(P,\bar{T}).$$

This measure indicates the difference between target and non-target sets. An obvious difference between them is expected for positive frequent impact-oriented exceptional behavior patterns. If $Cdr(P,|^{D}_{\bar{D}})$ is larger than a given threshold, then P far more frequently leads to positive than negative impact.

Definition 6. The *relative risk ratio* $Rrr(P,|^{T}_{\bar{T}})$ of P leading to target exceptional behavior classes D and non-target class $\bar{D}$ is defined as

$$Rrr(P,|^{T}_{\bar{T}}) = Prob(\mathrm{T}|P) / Prob(\bar{\mathrm{T}}|P) = Prob(P,T) / Prob(P,\bar{T})$$

$$= Supp_A(P,T) / Supp_A(P,\bar{T}).$$

This measure indicates the statistical difference of a sequence P leading to positive or negative impact in a global manner. An obvious difference between them is expected for positive frequent impact-targeted exceptional behavior patterns. In addition, if the statistical significance of P leading to T and $\bar{T}$ are compared in terms of local classes, then relative risk ratio $Rrr(P,|^{T}_{\bar{T}})$ indicates the difference of a pattern'*s* significance between target set and non-target set. If $Rrr(P,|^{T}_{\bar{T}})$ is larger than a given threshold, then P far more frequently leads to debt than results in non-debt.

Based on the above and other existing metrics such as *confidence*, *lift* and *Z-Score*, frequent impact-oriented exceptional behavior patterns can be studied to identify positive impact-oriented exceptional behavior patterns and negative impact-oriented exceptional behavior patterns.

3.2 Impact-Contrasted Exceptional Behavior Patterns

Difference between target activity set D and non-target set $\bar{D}$ may present useful contrast information in finding impact-targeted exceptional behavior patterns. For instance, exceptional behavior itemset P may satisfy one of the following scenarios:

- $Supp_D(P,T)$ is high but $Supp_{\bar{D}}(P,\bar{T})$ is low,
- $Supp_D(P,T)$ is low but $Supp_{\bar{D}}(P,\bar{T})$ is high.

In each of the above two cases, if there is a big contrast between two supports, say if $Supp_D(P,T)$ is much greater than $Supp_{\bar{D}}(P,\bar{T})$, it indicates that P is more or less associated with positive rather than negative impact, or vice versa.

In practice, those frequent itemsets P in D ({P--> T}) but not in $\bar{D}$ ({ $P \rightarrow \bar{T}$ }) are interesting because they tell us which exceptional behavior or exceptional behavior sequences lead to positive impact. In other cases, those frequent items in $\bar{D}$ ({ $P \rightarrow \bar{T}$ }) but not in D { $P \rightarrow \bar{T}$ } may help understand which activity sequences could prevent positive impact. Therefore, we define *impact-contrasted patterns* $P_{T\backslash\bar{T}}$ and $P_{\bar{T}\backslash T}$ as follows.

Definition 7. Given local frequent exceptional behavior itemset *P*, a *positive impact-contrasted pattern* $P_{T\backslash\bar{T}}$ exists if *P* is frequent in set *D* but not in set $\bar{D}$.

$$P_{T\backslash\bar{T}}: \{\, P \rightarrow T \,,\; P \rightarrow \bar{T} \,\}$$

Definition 8. Given local frequent activity itemset *P*, a *negative impact-contrasted pattern* $P_{\bar{T}\backslash T}$ exists if *P* is frequent in $\bar{D}$ but not in *D*.

$$P_{\bar{T}\backslash T}: \{\, P \rightarrow \bar{T} \,,\; P \rightarrow T \,\}.$$

After mining $P_{T\backslash\bar{T}}$, those itemsets with negative impact can be checked to see whether they trigger patterns {*P*--> *T*} or not. It is useful in applications where an exceptional behavior or exceptional behavior sequence leads to non-target impact. If yes, then they more likely lead to positive impact. $P_{\bar{T}\backslash T}$ represents frequent itemsets that are potentially interesting for non-target exceptional behavior.

Further, to measure the interestingness of frequent impact-contrasted exceptional behavior patterns, we define contrast supports and contrast lifts for $P_{T\backslash\bar{T}}$ and $P_{\bar{T}\backslash T}$, respectively.

Definition 9. Given a positive impact-contrasted exceptional behavior pattern $P_{T\backslash\bar{T}}$, the *positive contrast support* $CSupp_D(P_{T\backslash\bar{T}})$ and *positive contrast lift* $CLift_D(P_{T\backslash\bar{T}})$ are defined as follows. They tell us how much the lift of $P_{T\backslash\bar{T}}$ is.

$$CSupp_D(P_{T\backslash\bar{T}}) = Supp_D(P,T) - Supp_{\bar{D}}(P,\bar{T})$$

$$CLift_D(P_{T\backslash\bar{T}}) = Supp_D(P,T) / Supp_{\bar{D}}(P,\bar{T}) \;=\; Cdr(P, |_{\bar{D}}^{D})$$

Definition 10. Given a negative impact-contrasted pattern $P_{\bar{T}\backslash T}$, the *negative contrast support* $CSupp_{\bar{D}}(P_{\bar{T}\backslash T})$ and *negative contrast lift* $CLift_{\bar{D}}(P_{\bar{T}\backslash T})$ are defined as follows. They tell us how much the lift of $P_{\bar{T}\backslash T}$ is:

$$CSupp_{\bar{D}}(P_{\bar{T}\backslash T}) = Supp_{\bar{D}}(P,\bar{T}) - Supp_D(P,T)$$

$$CLift_{\bar{D}}(P_{\bar{T}\backslash T}) = Supp_{\bar{D}}(P,\bar{T}) / Supp_D(P,T) \;=\; Cdr^{-1}(P, |_{\bar{D}}^{D}) \cdot$$

4 Experiments

We tested [1] the above-discussed patterns on Centrelink debt-related activity data [3]. We used four data sources, *activity files* recording activity details, *debt files* containing debt details, *customer files* containing customer profiles, and *earnings files* storing earnings details. Our experiments analyzed activities related to both income and non-income related debts. To analyze the relationship between activity and debt, data from activity files and debt files was extracted. The timeline used in the activity data was between the 1st Jan and the 31st Mar 2006. We extracted 15,932,832 activity transactions recording government-customer contacts for 495,891 customers, leading to 30,546 debts in the first three months of 2006.

Based on the proposed activity construction strategy, we construct 454,934 sequences: 16,540 (3.6%) activity sequences associated with debts and 438,394 (96.4%) sequences with non-debts. These sequences contain 16,540 debts and 5,625,309 activities.

Table 2 shows frequent impact-oriented activity patterns discovered from the above unbalanced activity dataset. In the table, "LSUP" and "RSUP" denote the supports of a pattern's antecedent and consequent respectively. "CONF", "LIFT" and "ZSCORE" stand for the confidence, lift and z-score of the rule. From the table, we can see that the rule ("a_1, a_2 --> DET") has high confidence and lift. Its supports are very low, which are actually caused by the unbalanced class size (only 3.6% are activity sequences of debts). The second rule ("a_1 --> DET") is also of high lift (6.5), but the appearance of "a_2" triples the lift of the first rule.

Table 2. Frequent debt-oriented activity patterns discovered in unbalanced set

Frequent patterns	LSUP	RSUP	SUPP	CONF	LIFT	Z-SCORE
a_1 --> DET	0.0626	0.0364	0.0147	0.2347	6.5	175.7
a4 -> DET	0.1490	0.0364	0.0162	0.1089	3.0	99.3
a1, a4 -> DET	0.0200	0.0364	0.0125	0.6229	17.1	293.7
a_1, a_2 --> DET	0.0015	0.0364	0.0011	0.7040	19.4	92.1

Table 3 presents a sequential impact-contrasted pattern discovered in target and non-target data sets. The pattern pair, "a_4-->DET and a_4 --> NDT", has $CLift_D$=3.24, shows that a_4 is 2.24 times more likely to lead to debt than non-debt.

Table 3. Impact-contrasted activity pattern identified in target and non-target sets

Patterns (-->DET/NDT)	$Supp_D$	$Supp_{\bar{D}}$	$CSup_D$	$CLift_D$	$CSup_{\bar{D}}$	$CLift_{\bar{D}}$
a_4	0.446	0.138	0.309	3.24	-0.309	0.31
a_5	0.169	0.117	0.053	1.45	-0.053	0.69
a_4, a_5	0.335	0.107	0.227	3.12	-0.227	0.32
a_5, a_4, a_6,	0.241	0.077	0.164	3.13	-0.164	0.32

In this section, we illustrate some examples of high impact exceptional government-customer contact patterns identified in the Australian social-security activity data. The work was produced in close cooperation and on-spot assessment and iterative refinement by senior business analysts and managers in Centrelink. The Summary Report [4] delivered to the Centrelink Executives, the findings are deemed as very interesting to understand customer behavior, streamlining business processes, and preventing customer government debt.

5 Conclusions

High impact exceptional behavior is hard to be identified in massive data in which only rare and dispersed high impact behavior is of interest. The high impact exceptional behavior data presents special structural complexities, in particular, *unbalanced class* and *itemset distribution*. Mining rare exceptional behavior leading to significant impact to business is worthwhile data mining research.

In this paper, we have identified the following types of interesting impact-oriented exceptional behavior patterns in unbalanced activity data: (1) impact-oriented exceptional behavior patterns leading to either positive or negative impact, (2) impact-contrasted exceptional behavior patterns differentiating the significance of the same exceptional behavior resulting in contrast impact in target and non-target sets. New technical interestingness metrics have been developed for evaluating the above impact-targeted exceptional behavior patterns.

We have demonstrated the proposed impact-targeted activity patterns in analyzing Australian social-security activity data. The findings are of interest to Centrelink. The identified approach is also useful for analyzing exceptional behavior in many other applications, say national security and homeland security for counter-terrorism, distributed crimes and frauds, financial security, social security, intellectual property security etc.

Acknowledgement

Thanks are given to Ms Yvonne Morrow, Mr Rick Schurmann, Mr Peter Newbigin, Ms. Carol Ey, Ms. Michelle Holden at Centrelink Australia for their domain knowledge and data.

References

[1] Zhao, Y., Cao, L.: Full report: Improving income reporting (May 31, 2006)
[2] Cao, L., Zhang, C.: Domain-driven data mining: a practical methodology. Int. J. of Data Warehousing and Mining (2006)
[3] Centrelink. Integrated activity management developer guide, Technical Report, 30 (September 1999)
[4] Centrelink: Summary report: Improving income reporting (June 2006)
[5] Chen, H., Wang, F., Zeng, D.: Intelligence and security informatics for homeland security: information, communication, and transportation. IEEE Transactions on Intelligent Transportation Systems 5(4), 329–341 (2004)
[6] Wu, X., Zhang, C., Zhang, S.: Efficient Mining of Both Positive and Negative Association Rules. ACM Transactions on Information Systems 22(3), 381–405 (2004)
[7] Zhang, J., Bloedorn, E., Rosen, L., Venese, D.: [7] Zhang, J. In: Perner, P. (ed.) ICDM 2004. LNCS (LNAI), vol. 3275, pp. 571–574. Springer, Heidelberg (2004)
[8] Cao, L., Zhao, Y., Zhang, C., Zhang, H.: Activity Mining: from Activities to Actions. International Journal of Information Technology & Decision Making 7(2) (2008)
[9] Cao, L., Zhao, Y., Zhang, C.: Mining Impact-Targeted Activity Patterns in Imbalanced Data. IEEE Trans. on Knowledge and Data Engineering (to appear)