Rare Association Rule Mining

Petr Glos

Knowledge Discovery Lab

Faculty of Informatics

Masaryk University

glos@ics.muni.cz

Brno, October 18th 2011

Agenda

- References
- Introduction
- MSApriori
- RSAA
- Clustering for pre-procesing
- Temporal sequence associations
- Co-Location Patterns with Rare Events

- Questions

References

- Chandola V., Banerjee A. and Kumar V. Anomaly Detection: A Survey. ACM Computing Surveys, Vol. 41, No. 3, July 2009
- Hyunyoon Y., Danshim H.,Buhyun H.,Keun H. R. Mining association rules on significant rare data using relative support, The Journal of Systems and Software 67, Elsevier, 2003
- Koh Y. S., Pears R. Rare Association Rule Mining via Transaction Clustering, In Proc. Seventh Australasian Data Mining Conference (AusDM 2008), Glenelg, South Australia. CRPIT
- Chen J., He H., Williams G., Jin H. Temporal Sequence Associations for Rare Events, Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, 2004, Volume 3056/2004, 235-239
- Vilalta R., Ma S. Predicting Rare Events in Temporal Domains, Proccedings of the 2002 IEEE International Conference on Data Mining

Introduction

- Association Rule Mining
  - Rule Ant(ecendent) => Con(sequent) X => Y
  - **High** Support a/(a+b+c+d)
  - High Confidence a/(a+b)
  - Supp > minSupp => frequent itemsets
  - Conf > minConf => rules
  - Apriori algorithm – k-1 itemsets => k itemsets

	Suc	-Suc
Con	a	b
-Con	c	d

- **Rare** Association Rule Mining
  - **Low** Support
  - High Confidence
  - Supp < minSupp => rare itemsets
  - Conf > minConf => "rare" rules
  - Apriori algorithm extension or modification
  - Seeking frequent patterns with occurrences before rare events

Multiple Support Apriori Algorithm MSApriori

- Support depends on frequency of data items
- Minimum item support MIS for data item i

  $MIS(i) = MI(i)$ if $MI(i) > LS$

  $= LS$ otherwise
- $MI(i) = \beta * f(i)$
- $0 <= \beta <= 1$
- $f(i)$ data frequency
- LS least support

Relative Support Apriori Algorithm RSAA

- Significant rare data is one which its frequency in the database does not satisfy the minimum support but appears associated with the specific data in high proportion of its frequency.

- 1st support – used in process of frequent items discovery
- 2nd support / used in process of rare items discovery
- 1st support > 2nd support

- Relative support
  $\mathrm{Rsup}(i_1,\ldots i_k) = \max\{ \mathrm{sup}(i_1,\ldots i_k)/\mathrm{sup}(i_1) ,\ldots, \mathrm{sup}(i_1,\ldots i_k)/\mathrm{sup}(i_k) \}$

- Group of itemsets satisfied 1st support
- Group of itemsets not satisfied 1st but satisfied 2nd support
- Iteration process to generate "rare itemset" candidates

Rare Association Rule Mining via Transaction Clustering

- Pre-process by clustering transactions before performing association rule mining
  - Common set of large items – min support treshold
  - Seed Generation Phase - based on relative support
  - Allocation Phase – based on Jaccard similarity
- Apriori-Inverse on clusters generated
- minsup < sup(i) < maxsup

Temporal Sequence Associations for Rare Events
Predicting Rare Events in Temporal Domains

- Collection of entities ε_i ϵ E (i=1,…,n)
- Event sequence – s_i = { ($e_{i1,}$ t_{i1}) ,…, ($e_{ij,}$ t_{ij}),…, ($e_{in_i,}$ t_{in_i}) },

  (e_{ij} event type, t_{i1} timestamp)
- Target events T – events of given type from E
- Time window [t_s,t_e], constant length
- Windowed segment { ($e_{ip,}$ t_{ip}) ,…, ($e_{iq,}$ t_{iq}) },

  t_s <=t_{ip} <= t_{ip+1} <= … <= t_{iq} <=t_e
- Target segment – window segment with first occurrence time of target event
- Supp(p) in T
- Risk ratio
- Interesting patterns for target events
- Seeking frequent patterns for occurences of rare events

Mining Co-Location Patterns with Rare Events from Spatial Datasets

- Co-Location Pattern C – group of spatial feature/events that are frequently co-located in the same region.
- Spatial feature f is rare if its instances are substantially less than those of other features in a co-location.
- Participation ratio - Wherever the feature f is observed, with probability pr(C,f), all other features in C are also observed in neighbor-set.
- Participation index - Wherever any feature from C is observed, with probability of at least PI(C), all other features in C can be observed in neighbor-set.
-
- Seeking of Co-Location Patterns
- Modification of Apriori algorithm
- maxPrune algorithm

Questions ?

Thank you for your attention.