Applying Objective Interestingness Measures in Data Mining Systems

Robert J. Hilderman and Howard J. Hamilton

Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{hilder,hamilton}@cs.uregina.ca

Abstract. One of the most important steps in any knowledge discovery task is the interpretation and evaluation of discovered patterns. To address this problem, various techniques, such as the chi-square test for independence, have been suggested to reduce the number of patterns presented to the user and to focus attention on those that are truly statistically significant. However, when mining a large database, the number of patterns discovered can remain large even after adjusting significance thresholds to eliminate spurious patterns. What is needed, then, is an effective measure to further assist in the interpretation and evaluation step that ranks the interestingness of the remaining patterns prior to presenting them to the user. In this paper, we describe a two-step process for ranking the interestingness of discovered patterns that utilizes the chi-square test for independence in the first step and objective measures of interestingness in the second step. We show how this two-step process can be applied to ranking characterized/generalized association rules and data cubes.

1 Introduction

Techniques for finding association rules have been widely reported in the literature, commonly within the context of discovering buying patterns from retail sales transactions [1]. In the *share-confidence framework* [6], an *association rule* is an implication of the form $X \Rightarrow Y$, where X and Y are items (or sets of items), and the implication holds with *confidence* c and *share* s, if the number of items contained in X comprise $c\%$ of the number of items contained in $X \cup Y$, and the number of items contained in $X \cup Y$ comprises $s\%$ of the total number of items in the database. A *characterized itemset* is an itemset in which the corresponding transactions have been partitioned into classes based upon attributes which describe specific characteristics of the itemset [6]. A *generalized itemset* is one where the values of one or more characteristic attributes are generalized according to a taxonomic hierarchy [13].

As a result of the widespread adoption of on-line analytical processing, the study of data cubes is also receiving attention in the literature [4]. A *data cube*, also known as a summary table, is a redundant, multidimensional projection of a relation. Data cubes describe materialized aggregate views that group the

D.A. Zighed, J. Komorowski, and J. Zytkow (Eds.): PKDD 2000, LNAI 1910, pp. 432–439, 2000.

unconditioned data in the original database along various dimensions according to SQL groupby operators associated with measure attributes. Dimensions in data cubes can also be generalized according to taxonomic hierarchies [4].

The number of patterns generated by a knowledge discovery task can exceed the capacity of a user to analyze them efficiently and effectively, and this is a widely recognized problem. In response to this problem, various techniques/metrics have been proposed to prune those patterns that are most likely to be considered uninteresting or not useful. Many successful techniques utilize the chi-square test for independence [12] which is based upon the differences between the expected number of occurrences for a discovered attribute value combination and the observed number. The primary assumption is that these differences will be less for independent attributes. A chi-square value that has a low probability of occurring strictly by chance leads to a rejection of the hypothesis that the attributes are *independent*, and attributes that are not independent are considered to be *associated*. Other statistics, known as *measures of association* [3], can be used to determine the relative strength of discovered patterns.

Although the chi-square test and measures of association can be used to reduce the number of patterns that must be considered by a user, when mining a large database, the number of patterns discovered can remain large even after adjusting significance thresholds to eliminate spurious patterns. What is needed, then, is an effective measure to assist in the interpretation and evaluation step that ranks the interestingness of the remaining patterns prior to presenting them to the user.

In this paper, we describe a two-step process for ranking the interestingness of discovered patterns that utilizes the chi-square test for independence in the first step and objective measures of interestingness in the second step. We show how this two-step process can be applied to ranking characterized/generalized association rules and data cubes. We introduced this use of diversity measures for ranking discovered patterns in [7, 8]. We also identified five diversity measures, known as the *PHMI set* (i.e., principled heuristic measures of interestingness), that satisfy five principles of interestingness proposed in [9].

2 Background

The five diversity measures in the PHMI set are shown in Figure 1. These diversity measures consider the frequency or probability distribution of the values in some numeric measure attribute to assign a single real-valued index that represents the interestingness of a discovered pattern relative to other discovered patterns. Let m be the total number of values in the numeric measure attribute. Let n_i be the i-th value. Let $N = \sum_{i=1}^{m} n_i$ be the sum of the n_i's. Let p be the actual probability distribution of the n_i's. Let $p_i = n_i/N$ be the actual probability for t_i. Let q be a uniform probability distribution of the values. Let $\bar{q} = 1/m$ be the probability for t_i, for all $i = 1, 2, \ldots, m$ according to the uniform distribution q. For a thorough discussion of the PHMI set, see [7, 8].

$$I_{Variance} = \frac{\sum_{i=1}^{m} (p_i - \bar{q})^2}{m-1}$$

$$I_{Simpson} = \sum_{i=1}^{m} p_i^2$$

$$I_{Shannon} = -\sum_{i=1}^{m} p_i \log_2 p_i$$

$$I_{Total} = m * I_{Shannon}$$

$$I_{McIntosh} = \frac{N - \sqrt{\sum_{i=1}^{m} n_i^2}}{N - \sqrt{N}}$$

Figure 1. The PHMI set of diversity measures

3 Applications

In this section, we present two applications for objective measures of interestingness in data mining systems: ranking (1) characterized/generalized association rules and (2) data cubes. Due to space limitations, we do not describe the chi-square test for independence, as this topic is covered in other work [12], and we do not use it in the derived examples that follow. Instead, use of the the chi-square test for pruning results is demonstrated in the experimental results of the next section. Also, we do not dwell on techniques for generating characterized/generalized association rules and data cubes, as these techniques have also been covered in other work [1, 4, 7]. Here, we assume that these techniques are understood, at an intuitive level at least, and restrict our presentation to the use of diversity measures for ranking discovered patterns.

Input is provided by a sales database consisting of the *Transact* and *Cust* tables, shown in Tables 1 and 2, respectively. In the *Transact* table, the *TID* column describes the transaction identifier, the *LI* column describes a unique line item identifier within the corresponding transaction identifier, the *CID* column describes the identifier of the customer who initiated the transaction, the *Loc* column describes the location where the transaction was processed, the *ItemNo* column describes the item sold in the corresponding line item, the *Qty* column describes the quantity of the corresponding item that has been sold. In the *Cust* table, the *CID* column is the customer identifier, the *Loc* column describes the location where the customer lives, and the *Name* column describes the name corresponding to the customer identifier. The *Transact* and *Cust* tables can be joined on the *Transact.CID* and *Cust.CID* columns. The *Cust.Loc* column shown in the *Transact* table is a result of such a join, and is shown for reader convenience in the presentation that follows. The values in the *Transact.Loc* and *Cust.Loc* columns can be generalized according to the DGG shown in Figure 2.

3.1 Characterized/Generalized Association Rules

Using the characterized itemset generation algorithm, *CItemset* [5, 6], from the share-confidence framework, a minimum share threshold of 30%, a mini-

Table 1. The *Transact* table

TID	*LI*	*CID*	*Loc*	*ItemNo*	*Qty*	*Cust.Loc*
1	1	4	2	A	2	3
1	2	4	2	B	1	3
1	3	4	2	C	4	3
2	1	1	3	A	2	4
2	2	1	3	B	3	4
2	3	1	3	C	1	4
2	4	1	3	D	2	4
3	1	2	1	A	3	2
3	2	2	1	C	5	2
3	3	2	1	E	1	2
4	1	3	4	D	5	1
5	1	1	2	B	2	4
5	2	1	2	C	1	4
6	1	4	1	B	7	3
6	2	4	1	C	3	3
6	3	4	1	E	2	3
7	1	3	3	A	5	1
7	2	3	3	C	8	1
8	1	2	4	D	6	2
8	2	2	4	E	3	2
9	1	3	2	A	2	1
9	2	3	2	B	4	1
9	3	3	2	C	1	1
10	1	1	3	C	5	4
11	1	4	1	B	4	3
11	2	4	1	C	6	3
11	3	4	1	D	2	3
11	4	4	1	E	7	3
12	1	2	4	A	3	2
12	2	2	4	C	8	2
12	3	2	4	D	1	2
13	1	3	3	E	2	1

Table 2. The *Cust* Table

CID	*Loc*	*Name*
1	4	Smith
2	2	Jones
3	1	White
4	3	Black

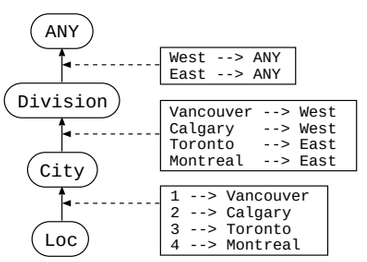

Figure 2. The *Loc* DGG

mum confidence threshold of 50%, the multi-attribute generalization algorithm, *All_Gen* [10, 11], and the SQL statement `SELECT TID, Cust.Loc, Qty FROM Transact, Cust WHERE Transact.CID = Cust.CID`, two of the many possible association rules and the corresponding summaries that can be generated according to the DGG in Figure 2, are shown in Table 3. In Table 3, the *Rule* ($x \Rightarrow y$) column describes the discovered association rule, the *Share* and *Conf.* columns describe the *global share* and *count confidence* of the corresponding association rule, respectively, as described in [5, 6], the *Node* column describes the level to which the values in the *Cust.Loc* column have been generalized according to the DGG in Figure 2, the *TIDs* column describes the transactions from the *Transact* table aggregated in each row as a result of generalizing the values in the *Cust.Loc* column (TIDs are not actually saved in practice), the *Cust.Loc* column describes the characteristic attribute, the *Qty(x)* and *Qty(y)* columns describe the *local item count*, as described in [5, 6], for the antecedent and consequent, respectively, of the corresponding association rule, the *Qty* ($x \cup y$) column describes the sum of *Qty(x)* and *Qty(y)*, and the *Count* column describes the number of transactions aggregated in each row.

Table 3 shows that association rule $C \Rightarrow A$ has share and confidence of 39.6% and 61.4%, respectively. Share is calculated as the sum of the quantity of all items in itemset $\{C, A\}$ divided by the quantity of all items in the *Transact* table (i.e., $44/111 = 39.6\%$). Confidence is calculated as the quantity of item C in itemset $\{C, A\}$ divided by the sum of the quantity of all items in itemset $\{C, A\}$ (i.e., $27/44 = 61.4\%$). The values of *Qty(C)* and *Qty(A)*, corresponding to

Table 3. Characterized/generalized association rules generated

Rule $(x \Rightarrow y)$	*Share (%)*	*Conf. (%)*	*Node*	*TIDs*	*Cust.Loc*	*Qty(x)*	*Qty(y)*	*Qty* $(x \cup y)$	*Count*
$C \Rightarrow A$	39.6	61.4	*City*	3, 12	*Calgary*	13	6	19	2
				7, 9	*Vancouver*	9	7	16	2
				1	*Toronto*	4	2	6	1
				2	*Montreal*	1	2	3	1
			Division	3, 7, 9, 12	*West*	22	13	35	4
				1, 2	*East*	5	4	9	2
$B \Rightarrow C$	33.3	56.8	*City*	2, 6, 11	*Montreal*	14	10	24	3
				5, 9	*Calgary*	6	2	8	2
				1	*Toronto*	1	4	5	1
			Division	1, 2, 6, 11	*East*	15	14	29	4
				5, 9	*West*	6	2	8	2

transactions 3 and 12, are 13 and 6, respectively, and calculated as the quantity of items C and A sold in the two transactions.

The values in the *Qty* $(x \cup y)$ and *Count* columns, called *vectors*, describe distributions of the quantity of items and the number of transactions, respectively, in the corresponding itemset, and these distributions can be used by the measures in the PHMI set to determine the relative interestingness of the corresponding association rule. For example, in Table 3, the vectors in the *Qty* $(x \cup y)$ column of the *City* and *Division* summaries for association rule $C \Rightarrow A$ are $(19, 16, 6, 3)$ and $(35, 9)$, respectively, and for association rule $B \Rightarrow C$ are $(24, 8, 5)$ and $(29, 8)$, respectively. The interestingness for these four summaries, according to $I_{Variance}$ (due to space limitations, we only discuss the results obtained for $I_{Variance}$) is $0.008815, 0.174587, 0.076211$, and 0.161066, respectively. Thus, the rank order of the four association rules (from most to least interesting) is $C \Rightarrow A$ (*Division*), $B \Rightarrow C$ (*Division*), $B \Rightarrow C$ (*City*), and $C \Rightarrow A$ (*City*).

3.2 Data Cubes

Using the multi-attribute generalization algorithm, *All_Gen* [10, 11], and the SQL statement `CREATE VIEW ItemsByLoc (Transact.Loc, Item, Cust.Loc, TotalQty) AS SELECT Transact.Loc, Item, Cust.Loc, SUM (Qty)) AS TotalQty FROM Transact, Cust WHERE Transact.CID = Cust.CID GROUPBY Item, Transact.Loc, Cust.Loc`, four of the eight data cubes that can be generated according to the DGG in Figure 2, are shown in Figure 3. Figure 3 actually describes four data cubes because each cell contains two values; the top value is the quantity of items aggregated from the transactions, and the bottom value is the number of transactions aggregated. The *Item* attribute is on the vertical dimension, *Transact.Loc* on the horizontal, and *Cust.Loc* on the diagonal. The *City* and *Division* labels describe the level to which the values in both the *Transact.Loc* and *Cust.Loc* dimensions have been generalized according to the DGG in Figure 2. The other four possible data cubes (not shown) are obtained by generalizing only one of the *Transact.Loc* and *Cust.Loc* dimensions, respectively, in each cube.

Within the context of data cubes, objective interestingness measures can be

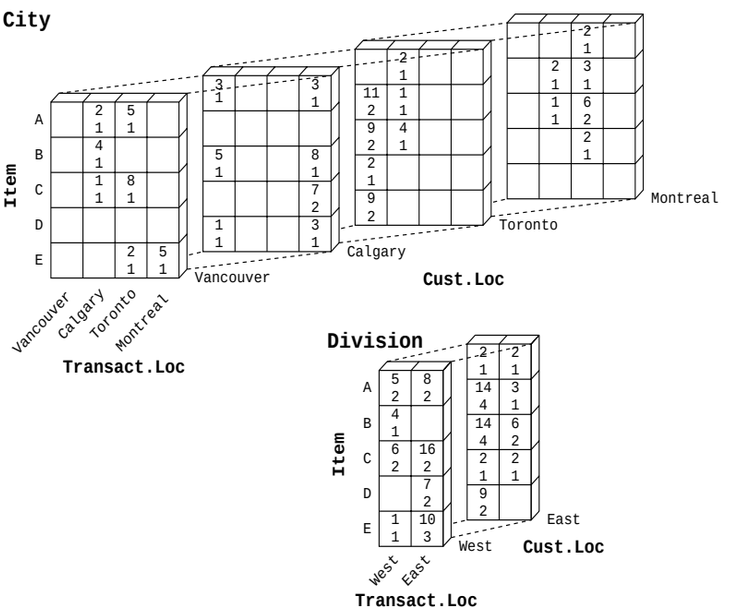

Figure 3. Data cubes generated

applied in three ways: (1) to the whole data cube, (2) to slices, and (3) to rows and columns (rows and columns example not shown).

Whole Data Cube. The *Qty* vectors for the *City* and *Division* data cubes are $(11, 9, 9, 8, 8, 7, 6, 5, 5, 4, 4, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1)$ and $(16, 14, 14, 10, 9, 8, 7, 6, 6, 5, 4, 3, 2, 2, 2, 2, 1)$, respectively. The interestingness of the *City* and *Division* data cubes, according to $I_{Variance}$, is 0.000761 and 0.001807, respectively. Thus, the rank order of the two data cubes is *Division* and *City*.

Slices. The *Qty* vectors for the *Vancouver*, *Calgary*, *Toronto*, and *Montreal* slices in the *Cust.Loc* dimension of the *City* data cube are $(8, 5, 5, 4, 2, 2, 1)$, $(8, 7, 5, 3, 3, 3, 1)$, $(11, 9, 9, 4, 2, 2, 1)$, and $(6, 3, 2, 2, 2, 1)$, respectively. The interestingness of the four slices, according to $I_{Variance}$, is $0.007969, 0.006931, 0.011740$, and 0.011879, respectively. Thus, the rank order of the four slices is *Montreal*, *Toronto*, *Vancouver*, and *Calgary*.

4 Experimental Results

A series of experiments were run using *DGG-Interest*, an extension to *DB-Discover*, a research data mining tool developed at the University of Regina [2].

DGG-Interest evaluates the summaries generated by *DB-Discover* using the proposed two-step process (again, we present only the $I_{Variance}$ results).

Input data was supplied by the NSERC Research Awards and Customer databases [7, 10, 11]. Summary results for six representative discovery tasks, where two to four attributes have been selected for discovery, are shown in Table 4. In Table 4, the *Task* column describes a unique discovery task identifier, the *Attributes* column describes the number of attributes selected, the *Generated* column describes the number of summaries generated by the corresponding discovery task, the *Pruned* (*%Pruned*) column describes the number (percentage) of summaries in which no significant association between attributes was found in the first step, and the *Associated* (*%Associated*) column describes the number (percentage) of summaries in which a significant association was found in the first step, and which are available for ranking in the second step. For example, in *N-2*, an NSERC discovery task, two attributes were selected, 22 summaries were generated, 14 (63.6%) were pruned, and a significant association was discovered between attributes in the remaining eight (36.3%), which were available for ranking in the second step.

Table 4. Summary results for seven representative discovery tasks

Task	*Attributes*	*Generated*	*Pruned*	*%Pruned*	*Associated*	*%Associated*
N-2	2	22	14	63.6	8	36.3
N-3	3	70	43	61.4	27	38.6
N-4	4	186	143	76.9	43	23.1
C-2	2	340	325	95.6	15	4.4
C-3	3	3468	3288	94.8	180	5.2
C-4	4	27744	26163	94.3	1581	5.7

Detailed results for the *N-2* discovery task are shown in Table 5. In Table 5, the *Summary* column describes a unique summary identifier, the *Tuples* column describes the number of tuples in each summary, the *Attributes* column describes the number of attributes containing more than one unique domain value, the χ^2-*Status* column describes the result of the chi-square test, the χ^2-*Value* column describes the calculated chi-square value, the *DF* column describes the degrees of freedom for the chi-square test, the $I_{Variance}$ and *Rank* columns describe the calculated interestingness and rank determined by $I_{Variance}$ after pruning those containing no significant associations. In the chi-square calculation, any zeroes occurring in the work contingency table associated with each summary are considered to be structural zeroes.

References

1. T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions: A case study. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 254–260, San Diego, California, August 1999.

Table 5. Detailed results for the N-2 discovery task

Summary	*Tuples*	*Attributes*	χ^2*-Status*	χ^2*-Value*	*DF*	$I_{Variance}$	*Rank*
1	5	2	Not Associated	0.343263	2	0.079693	-
2	11	2	**Associated**	11.343847	5	0.013534	1.0
3	10	2	Not Applicable	-	-	0.041606	-
4	21	2	**Associated**	33.401486	18	0.011547	4.0
5	50	2	**Associated**	77.999924	45	0.002078	6.0
6	3	1	Not Applicable	-	-	0.128641	-
7	6	1	Not Applicable	-	-	0.018374	-
8	4	1	Not Applicable	-	-	0.208346	-
9	4	2	Not Applicable	-	-	0.208346	-
10	9	2	Not Associated	7.003831	6	0.050770	-
11	21	2	Not Associated	23.186095	15	0.008896	-
12	10	1	Not Applicable	-	-	0.041606	-
13	5	1	Not Applicable	-	-	0.024569	-
14	9	1	Not Applicable	-	-	0.017788	-
15	2	1	Not Applicable	-	-	0.377595	-
16	2	2	Not Applicable	-	-	0.377595	-
17	9	2	Not Associated	6.731139	4	0.018715	-
18	16	2	**Associated**	12.431117	8	0.010611	2.0
19	40	2	**Associated**	63.910313	36	0.002986	5.0
20	67	2	**Associated**	97.799623	72	0.001582	7.0
21	17	2	**Associated**	18.221498	12	0.012575	3.0
22	30	2	Not Associated	24.920839	24	0.006470	-

2. C.L. Carter and H.J. Hamilton. Efficient attribute-oriented algorithms for knowledge discovery from large databases. *IEEE Transactions on Knowledge and Data Engineering*, 10(2):193–208, March/April 1998.
3. L.A. Goodman and W.H. Kruskal. *Measures of Association for Cross Classifications*. Springer-Verlag, 1979.
4. J. Han, W. Ging, and Y. Yin. Mining segment-wise periodic patterns in time-related databases. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 214–218, New York, New York, August 1998.
5. R.J. Hilderman, C.L. Carter, H.J. Hamilton, and N. Cercone. Mining association rules from market basket data using share measures and characterized itemsets. *International Journal on Artificial Intelligence Tools*, 7(2):189–220, June 1998.
6. R.J. Hilderman, C.L. Carter, H.J. Hamilton, and N. Cercone. Mining market basket data using share measures and characterized itemsets. In X. Wu, R. Kotagiri, and K. Korb, editors, *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 159–173, Melbourne, Australia, April 1998.
7. R.J. Hilderman and H.J. Hamilton. Heuristic measures of interestingness. In J. Zytkow and J. Rauch, editors, *Proceedings of the Third European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'99)*, pages 232–241, Prague, Czech Republic, September 1999.
8. R.J. Hilderman and H.J. Hamilton. Heuristics for ranking the interestingness of discovered knowledge. In N. Zhong and L. Zhou, editors, *Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, pages 204–209, Beijing, China, April 1999.
9. R.J. Hilderman and H.J. Hamilton. Principles for mining summaries: Theorems and proofs. Technical Report CS 00-01, Department of Computer Science, University of Regina, February 2000. Online at http://www.cs.uregina.ca/research/Techreport/0001.ps.
10. R.J. Hilderman, H.J. Hamilton, and N. Cercone. Data mining in large databases using domain generalization graphs. *Journal of Intelligent Information Systems*, 13(3):195–234, November 1999.
11. R.J. Hilderman, H.J. Hamilton, R.J. Kowalchuk, and N. Cercone. Parallel knowledge discovery using domain generalization graphs. In J. Komorowski and J. Zytkow, editors, *Proceedings of the First European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'97)*, pages 25–35, Trondheim, Norway, June 1997.
12. B. Liu, W Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 125–134, San Diego, California, August 1999.
13. R. Srikant and R. Agrawal. Mining generalized association rules. In *Proceedings of the 21th International Conference on Very Large Databases (VLDB'95)*, pages 407–419, Zurich, Switzerland, September 1995.