Identification of Multidimensional Relationship among Item Sets using Association Rules

Mamta
Shobhit University Meerut

Shwetank Arya
Gurukul Kangri University Hardwar

R. P. Agarwal
Shobhit University Meerut

ABSTRACT
In recent days, there is more interest in discovering multi dimensional relationship among item sets rather than frequent pattern sets for applications in specific domains viz. identification of irregularities in stock marketing, assessing the causes in certain diseases, identifying irregularities in farming system etc. This paper focuses on the mining of multi dimensional relationship among various item sets. An efficient algorithm to identify multi dimensional relationship in Inter Disciplined Independent Variables (IDIV) and dependent variable has been proposed. The effectiveness of the algorithm has been assessed on real world data set related to socio-economic conditions of farming system.

General Terms
Data mining task discovering multidimensional relationship among item sets.

Keyword
Multidimensional relationship, association rule, support, confidence.

1. INTRODUCTION
Association rule mining is an important data mining task involving the discovery of hidden relationships in items stored in huge repository. Various extensions of the traditional associations rule mining have been proposed so far, however the problem of mining of multi dimensional associations are not tackled yet to the satisfaction of the user. Multi dimensional associations are useful in many applications such as cross marketing, catalog design, predicting the failure of telecommunication switches, weather forecasting, minimizing socio-economic risk factors associated with farming system and many more. The availability of one factor when another factor is presents represents association rule. The importance of association rules is based on support and confidence.

Support(s) and Confidence(c)
A general form of an association rule is:

Body → Head [Support, Confidence]

Support of the rule A=>B:

Describe the frequency of the rule within all transactions in the database i.e., the probability that a transaction contain both A and B.
Support (A => B [s, c]) = P (A U B) = Support ({A, B})

Confidence of the rule A => B:

Describes the percentage of transaction containing A which also contain B.

Confidence (A = > B [s, c]) = P (B\A)
=P (AUB)/P (A)
=Support ({A, B})/Support ({A})

Minimum support threshold is a threshold for support, if it is:

High: valid rules which occurs often.

Low: valid rules which occurs rarely.

Minimum confidence threshold is a threshold for confidence, if it is:

High: few rules, but all 'almost logically true'

Low: Many rules, many of which are rarely uncertain.

Rules that satisfy both minimum support and minimum confidence are called strong rule. For instance, the information that the farmer has passed 12th class may be very innovative farmer. This can be represented in the following ways:

(X"12th-passed") => (X"innovative") [s=10%, c=80%]

A support 10% for association rule means that 10% of all the transactions under analysis show that 12th passed education and innovative attitude occur together in data set. The confidence or strength for an association rule X=>Y is the ratio of the number of transactions that contain X U Y to the number of transactions that contain X. The confidence 80% indicates that if farmer is 12th class passed, indicated in 10 transactions, then in 8 transactions out of 10 transactions he possessed innovative attitude.

Data repositories consist of large set of items and relationships but interested associations are few, which are measured by support and confidence. Several studies are available on frequent item sets and association rules. In this paper mining multi dimensional association rules has been discussed.

2. RELATED WORK
Researchers describe the data mining applications capable to help domain experts to integrate their knowledge into data transformation to generate a variety of possible patterns. For more complex data with mutual relationships, the derived patterns will be more complex and more valuable [5]. Researchers discussed the methods for mining frequent patterns and applications of frequent patterns and mentioned the need of a mechanism that provide the deep understanding and interpretation of complex patterns [6]. In another study a probability based evaluation metric was proposed and a mining algorithm was given to mine mutually exclusive items in transaction databases [7]. In a study various binary decision diagrams were used for solving pattern mining problems in a variety of situations such as frequent item sets, frequent subsequences and contrast mining. Binary decision diagram was found useful in mining item based patterns such as

frequent item sets and contrast mining [8]. Researchers identified mutually dependent patterns in computer networks, and conclude that the interrelated components are impacted by the same failure and strong mutual dependencies are common in computer networks [9-10]. In another study a general framework for assessment of similarity between both simple and complex patterns was explained. The similarity between two simple patterns of the same type was computed by combining, by means of an aggregation function [11]. In a study multi dimensional association rules are constructed and used with preprocessing method to reduce the mining time[12].

Researchers in all these studies discussed the need of a system that allow the user to determine more advanced and understandable pattern mining, which can uncover the hidden and valuable patterns and also discover intra dimensional and multi dimensional relationships in item sets. In the present study single dimensional and multi dimensional rules are discussed with the example of socio-economic factors affecting farming system.

2.1 Multi Dimensional Association Rule

Association rule which involve single predicate is referred to as single dimensional or intra dimensional association rule [2]. For instance, in mining database containing socio-economic factors of farmers, we may discover Boolean association rule in the following way:

Buy (X,"Seed") => Buy (X, "Tractor")

In some application data set is multi dimensional. For instance in addition to store the information of farmers' income per annum, data set also record the attributes associated with income such as age, education, extension-use-services, live stock, experience, innovativeness, risk willingness, land size etc. Considering each database attribute or warehouse dimension as a predicate, we can mine association rules containing multiple predicates, such as:

age (X, "25….75") ^ education (X, "5^{th}….12^{th}) => income (X, "100 thousands…….,999 thousands"). (1.1)

Education (X."5^{th}.12^{th}) ^ extension-services (X,"1…4") ^innovative (X,"1…4") => income (X,"1lac …100lac". (1.2)

Association rules that involve two or more dimensions can be referred to as multidimensional association rules. Rule 1.1 contains three predicates age, education and income, each of which occurs only once in the rule. It has no repeated predicate. Multidimensional association rule with no repeated predicates are called inter dimensional association rule [2]. Mining for multidimensional patterns could be very useful in discovering inherent dependency of one factor on another factor. For instance, a farmer may get higher income due to the higher education, experience, bigger land area, more live stock, more use of extension services, more innovative attitude, more risk willingness or any other associated factor. After indentifying the important determinant associated with farming, that determinant can be improved to improve the productivity consequently income of the farmer can be improved. Present study uses the example of socio-economic factors related to farming conditions. Ten factors are Inter Disciplined Independent Variables (IDIV) factors and one is dependent factor that is farmers' income per annum. In this paper an algorithm is designed and discussed to identify intra dimensional and multi dimensional patterns from the large repository of socio-economic factors of farmers affecting farming system [3-4]. Present study uses one dependent variable income and 10 inter disciplined independent variables which are structured using frame fig.1.

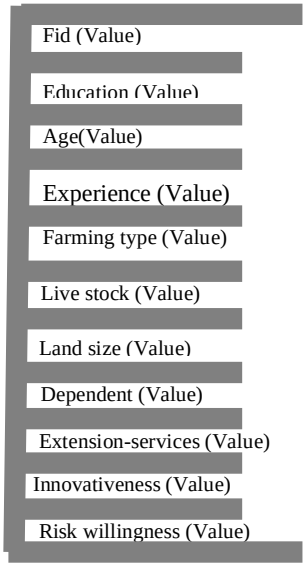

Fig1. Frame structure to display Inter disciplined independent variable.

Following algorithm is proposed to mine intra dimensional and multi dimensional association rules.

```
Algorithm
multidimensional_relationship(code1,code2,thresholdValue1,t
hresholdValue2)

income:farmers' income per annum
code1: IDIV1
code2: IDIV2
support: support factor
averageIncome: farmers' average income
total: temporary variable to add income of farmer
count: to store support factor
Begin
{
while not eof()
{
if( code1==thresholdValue1 and code==thresholdValue2)
{
total=total+income
count++
}
}
display multidimensional relationships in code1 and code2
corresponding to threshold value1 and threshold value2
display corresponding support factor
display average income
display confidence after calculating the confidence
```

Fig. 2. The algorithm scan the data set for count of each candidate, find the corresponding confidence and income of the farmer-fid.

3. EXPERIMENT

The proposed algorithm has been assessed on real world data set related to socio-economic conditions of farmers affecting farming system. Data set consisting of 11 variables has been collected from 325 farmers living in Jalalpur, Hapur and Siwaya villages located in district Meerut. Threshold value 1 to 4 is used as input to find the multi dimensional relationships in IDIV and dependent variable. For the assessment, following pairs of IDIV and dependent variable are considered for intra dimensional relationships. Results are depicted in graphical form from figure 3 to figure 6.

1. Dependents and income
2. Live stock and income
3. Land size and income
4. Innovativeness and income
5. Extension-services and income

4. RESULTS

Proposed algorithm identifies the intra dimensional relationships in IDIV and dependent variable income and provides the results which can be used to take the appropriate decisions to improve the farming system.

4.1 Dependent v/s Income

Graph shown in figure 4, indicates the relationships between family dependents and income and conclude that farmers' income increases with the increment in family dependents.

Table 1. Dependent variable, Income and support (%)

Dependent	Income(Th.)	Support (%)
1	210	3
2	172	44
3	202	36
4	277	16

Fig. 3.Dependents v/s Income (Threshold values as per table 6.)

4.2 Live stock v/s Income

Graph shown in figure 5, indicates the relationships between live stock and income, and conclude that farmers' income decreases as they increase their live stock.

Table 2. Live stock, Income and support (%)

Live stock	Income (Th.)	Support (%)
1	124	25
2	220	56
3	245	14
4	272	2

Fig. 4.Live stock v/s Income (Threshold values as per table 6.)

4.3 Land size v/s Income

Graph shown in figure 6, indicates the relationships between land size and income and conclude that farmers' income decrease as their land size increase.

Table 3. Land size, income and support (%)

Land size	Income	Support (%)
1	142	13
2	167	46
3	183	15
4	307	25

Fig. 5.Land size v/s Income (Threshold values as per table 6)

4.4 Extension services v/s income

Graph shown in figure 9, indicates the relationships between extension services and income and conclude that farmers' income increase when they use extension services regularly.

Table 4. Extension-services, Income and Support (%)

Extension-services	Income (Th.)	Support (%)
1	233	9
2	208	48
3	181	39
4	250	3

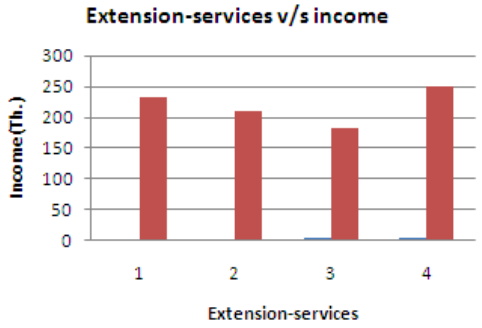

Fig. 6.Extension services v/s Income (Threshold values as per table 6)

Similarly all the pairs of IDIV and dependent variable can assessed to fine intra dimensional relationships.

4.5 Multi dimensional patterns

Proposed algorithm has been assessed on education and extension-services and dependent variable income to find multidimensional relationships.

Following table shows the threshold value of education, extension-services, income, support and confidence, assessed by the algorithm. Values having the confidence greater than and equal to minimum confidence (15%) are given in the table 5.

Table 5. Education, Extension-services and income per annum of farmers

Education	Extension-services	Income (Th.)	Support (%)	Confidence (%)
1	2	323	4	88
2	1	281	4	17
2	2	256	8	39
2	3	239	9	41
3	2	178	15	44
3	3	172	12	37
4	2	180	19	50
4	3	148	16	43

Table 6. Threshold value for Inter Disciplined Independent Variables (IDIV).

IDIV	Category	Threshold value
Age	AGE <=25	1
	Age>25&Age<=50	2
	Age>50&Age<=75	3
	Age>75	4
Education	Education < 8th class	1
	Education>=8&Education<10	2
	Education>=10&Education<12	3
	Education>12	4
Experience	Experience <10 years	1
	Experience >=10 & Experience <20	2
	Experience >=20 & Experience <30	3
	Experience >=30	4
Dependent	Dependents < 3	1
	Dependents >=3 & Dependents < 6	2
	Dependents >= 6 & Dependents < 8	3
	Dependents >=8	4
Live stock	Livestock < 3	1
	Livestock >=3 & Livestock < 6	2
	Livestock >= 6 & Livestock < 8	3
	Livestock >=8	4
Land size	Land size < 10	1
	Land size >=10 & Land size < 20	2
	Land size >=20 & Land size < 30	3
	Land size >=30	4
Extension-services(ES)	Never used extension-services	1
	ES >=1 & less than 4 times	2
	ES >=4 & less than 7 times	3
	ES >=7 times	4
Innovative	No innovative	1
	Innovative	2
	More Innovative	3
	Most Innovative	4
Risk Will	No risk willingness	1
	Risk willingness	2
	More risk willingness	3
	Most risk willingness	4

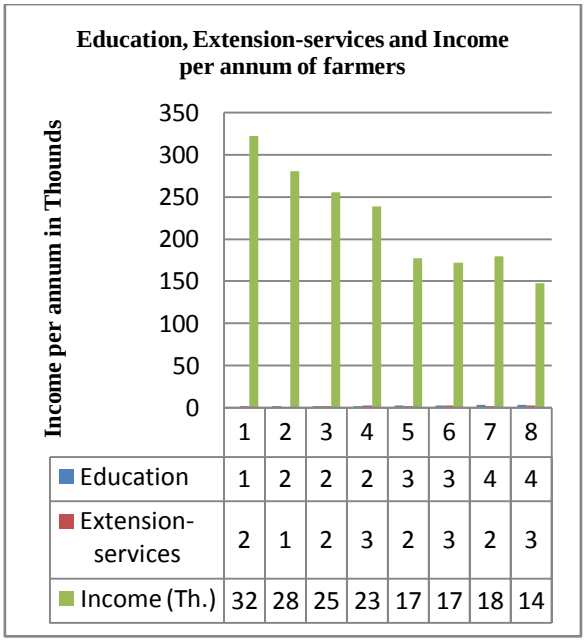

Fig. 7.(Education, Extension Services & Income, Threshold values as per table 6.)

The results conclude that the farmers' income is greatly influenced by education level and extension services. X axis shows threshold values of education and extension-services, Y axis indicate the income of farmers in thousands. The farmers having the education level of 12th class who visited extension-services1 to 4 times in a year, have 180 thousand income per annum with confidence factor of 50% and the farmers who visited extension-services 4 to 7 times in a year have income 180 thousand per annum with confidence factor 43%. Similar attitude is seen in those farmers whose education level is 10th class. It indicate that education facilitate the farmers to attend and understand extension-services program which contribute a lot in agricultural productivity consequently their income rise. Further, these programs conducted by private or government organizations must be so designed and delivered, that all farmers can use them.

Similarly, impact of more than two IDIV can be assessed on income of framers.

5. CONCLUSION

The association rule mining involves searching for relationships among item sets. These relationships can be further analyzed and may surface causal relationships and behavior. In this paper, an algorithm to find the multi dimensional relationship in IDIV affecting farmers' income per annum is given and assessed using real world dataset related to farming system. Results conclude that income is affected by various factors such as family dependents, live stock, extension-services and innovative attitude of the farmer. Result shown in the graphs depict that increment in the live stock has positive impact on the income. Regular visit of extension-services improve the farming and increase the income. Innovative attitude of the farmer give positive impact on the income. Similar relationships are found in family dependents and land size.

6. REFERENCES

[1] Dunham M.H., "Data Mining Introductory and Advanced Topics", Pearson Education,Edition 6th , 2009.

[2] Kamber M. and Han J. "Data Mining Concepts and Techniques", Edition 3rd , 2010

[3] Rajshree M. and Arya S., "Role of Data Mining in Minimizing Socio-Economic Risk Factors(SCRF) Affecting Agriculture", International Journal of Advanced Research in Computer Science, Volume 2, No. 5, Sept-Oct 2011.

[4] Rajshree M, Arya S. and Agarwal R.P., "Data Mining Techniques for Agriculture and Related Areas", International Journal of Advanced Research in Computer Science, Volume 2, No. 6, Nov-Dec 2011.

[5] Kriegel H.P., Karsten M., Borgwardt, P.K., Pryakhin Alexey, Schubert M. and Zimek A.,"Future trends in data mining", Data Min Knowl Disc., 2007.

[6] Han J., Cheng H., Xin D. and Yan X., "Frequent patterns: current status and future directions", Data Min Know Disc, 2007.

[7] Tzanis G. and Berheridis C.,"Mining for Mutually Exclusive Items in Transaction Databases", http://lpis.csd.auth.gr/publications/Tzanis_IJDWM07.pdf

[8] l.Elsa, "Mining Simple and Complex patterns efficiently using binary decision diagram", www.Biblioteca.Net, June 2009.

[9] Ma S. and Hellerstein J.L.,"Mining Mutually Dependent Patterns", IEEE,2001

[10] Ma S. and Hellerstein J.L.,"Mining Mutually Dependent Patterns for System Management", IEEE Journal on selected areas in communications,vol.20, No.4,May 2002.

[11] Bartolini I., Ciaccia P., Ntoutsi I, Patella M. and Theodoridis Y.,"A unified and Flexible Framework for comparing simple and complex patterns, Patterns for Next Generation Database System, 2004.

[12] Sug H.,"Discovery of Multidimensional Association rules Focusing on instances of Specific class", International Journal of Mathematics and Computers in Simulations, www.naun.org/journals/mcs/20-651.**pdf**

[13] Jadav J.J and Panchal M., "Association rule mining method on OLAP cube", International Journal of Engineering Research and Applications, Volume. 2, Issue 2, 2012.

[14] Nair B. and Tripathy A.K., "Accelerating Closed Frequent Itemset Mining by Elimination in Computing and Information Services, Volume 2. No. 2, 2011.