A Novel Rule based Data Mining Mechanism for Identification of Mutual Relationship among Item Sets

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 mutually dependent 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 mutual relationship among various item sets. An efficient algorithm to identify mutual relationship in Inter Disciplined Independent Variables (IDIV) 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 for identification of mutual relationship among item sets.

Keywords
Mining Mutual relationship, Rule based data mining, Socio-economic conditions, IDIV.

1. INTRODUCTION
Data mining is the process of identifying potentially useful, understandable and hidden patterns in large data repository. Advanced high throughput experimental technologies, high speed communication systems and internet facilities generate large volume of data repository automatically. Scientist, researchers and industrialist are facing difficulties in dealing with huge dataset which is too large for manual analysis. Data mining is the technique for discovering hidden and useful information from large data sets automatically. It is a new discipline of computer science, also referred as knowledge discovery. Fayyad et.al., describes knowledge discovery as searching of patterns [5]. Different methods of data mining are available such as classification, clustering and statistical methods. Various classification techniques are used to mine hidden and interesting patterns. Association rule mining is a popular classification technique to discover related and mutually dependent item sets. Several forms of association rule mining have existed however; the problem of mining for mutually dependent patterns has not been tackled so far [6].

Mining for mutually related item sets could be very useful in discovering inherent taxonomical information in a variety of situations. For instance, a farmer with higher income level may be with bigger land area and consequently he possesses comparatively more live stocks than the farmers with smaller land area. If the farmer possesses bigger land size than he may not be have low income level. It seems that the two factors exclude each other, If one item is present than the probability of the presence of other item may be less. Its vice versa is also true, if one factor is present than the probability of the presence of other factor may be higher. All these observation motivated to propose an algorithm to mine mutual relationship among item sets. These mutually related item sets play important role in discovering the answer of complex queries and making decisions. Present study uses the example related to the socio-economic conditions of farmers. 10 Inter Disciplined Independent Variables (IDIV) and one dependent variable are identified through the survey of 136 farmers. In this paper an algorithm based on association rules is designed and discussed to identify mutual relationship in IDIV affecting the income of farmers [9-10].

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 [1]. Researcher discussed the need of applications which can mine the complex patterns with mutual relationships within mined data sets but not suggested any mechanism or algorithm to perform the task. 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 [2]. 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 [3]. 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 [4].In another study a probability based evaluation metric was proposed and a mining algorithm was given to mine mutually exclusive items in transaction databases.[6]. 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 [7-8].

Researchers in all these studies mentioned the need of a mechanism that allow the user to determine the more complex, understandable and advanced patterns mining, which can reveal the hidden valuable patterns, interpret the patterns and also discover the mutual relationships among various items to help the domain expert in decision making.

3. MECHANISM FOR IDENTIFICATION OF MUTUAL RELATIONSHIP AMONG INTER DISCIPLINED INDEPENDENT VARIABLES

Mechanism to find the mutual relationship in various items can be described with the help of typical example of socio-economic conditions of farmers as shown in fig. 1.

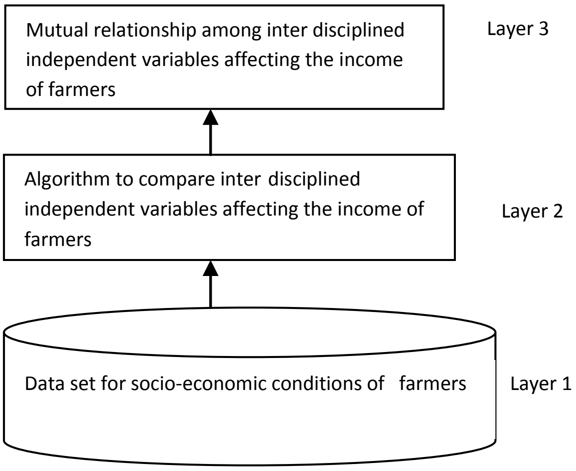

Fig 1: Mechanism of mining from data set.

The mechanism has been designed into layers. First layer is related with the repository of data set, which consists of dataset related to socio-economic conditions of farmers. Repository consists of threshold value of ten socio-economic inter disciplined independent variables affecting farmers' income per annum. Income is given in thousands. Threshold value of ten items is given in figure 3. Second layer uses the algorithm to compare IDIV. Top layer results the mutual relationship in various items. 10 IDIV and income related to socio-economic conditions of farmers are given in the fig. 2.

1 Age	2 Education
3 Live stock	4 Family dependants
5 Extension-services	6 Innovativeness.
7 Experience	8 Land Size
9 Self/Leased Land	10 Risk willingness
11 Income per annum.	

Fig 2:IDIV & income.

Threshold value 1-4 is used for the different category of each IDIV as shown in Fig. 3.

IDIV	Rule	Threshold value
Age	1.Age <= 25	1
	2. Age >25 & Age <=50	2
	3. Age >50 & Age <=75	3
	4. Age > 75	4
Education	1. Education < 8th class	1
	2. Education >=8 & Education <10	2
	3. Education >=10 & Education<12	3
	4. Education >=12	4
Experience	1. Experience <10 years	1
	2. Experience >=10 & Experience <20	2
	3. Experience >=20 & Experience <30	3
	4. Experience >=30	4
Dependents	1.Dependenst < 3	1
	2. Dependents >=3 & Dependents < 6	2
	3. Dependents >= 6 & Dependents < 8	3
	4. Dependents >=8	4
Live stock	1. Livestock < 3	1
	2. Livestock >=3 & Livestock < 6	2
	3. Livestock >= 6 & Livestock < 8	3
	4. Livestock >=8	4
Land size in Bigha	1. Land size < 10	1
	2. Land size >=10 & Land size < 20	2
	3. Land size >=20 & Land size < 30	3
	4. Land size >=30	4
Self/Lease land	1. Self	1
	2. Lease	2
Extension services (ES)	1. Never used ES	1
	2. ES >=1 & less than 4 times	2
	3. ES >=4 & less than 7 times	3
	4. ES >=7 times	4
Innovativeness	1. No innovative	1
	2. Innovative	2
	3. More Innovative	3
	4. Most Innovative	4
Risk Willingness	1. No risk willingness	1
	2. Risk willingness	2
	3. More risk willingness	3
	4. Most risk willingness	4

Fig 3: Threshold value for IDIV.

To mine the mutual relationship in item sets an algorithm is proposed as shown in fig. 4.

```
Algorithm compares (code1, code2, choice)

set count to 0;
code1: independent factor code
code 2:dependent factor code
choice: class of independent factor
total_value:variable to store total of dependent factor
begin
{
while(not eof())
{
        if (choice==code1)
        {
                total_value=total_value+code2;
                Count++;
        }
}
display mutual relationship in code1 and code2
average=total_value/count;
display average for code2;
display Count;
}
end.
```

Fig 4: Algorithm to identify mutual relationship in IDIV

3.1 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 136 farmers living in Jalalpur village located near to Modipuram in district Meerut. The goal of the experiment is to examine if the produced unified mechanism are semantically well formed. Threshold value 1 to 4 is used as input to find the mutual relationships in IDIV. Support factor, determinant of importance of association rule is also calculated and displayed in the tabulated form. For the assessment, following pairs of IDIV are considered and results are depicted in the graphical form from figure 5 to 14.

1. Age and live stock
2. Age and risk willingness
3. Education and risk willingness
4. Experience and live stock
5. Experience and extension-services
6. Experience and risk willingness
7. Dependent and live stock
8. Dependent and innovativeness
9. Dependent and extension-services
10. Live stock and land size
11. Innovativeness & Extension-services

3.2 Results

Graph shown in figure 5, indicates the mutual relationships between age and live stock and shows that 3% farmers having age group of 50 to 75 years possessed more than 8 live stocks while the farmers having age less than 50 years possessed less live stocks.

Age	Live stock	Support (%)
2	1	40
2	2	33
2	3	22
2	4	3

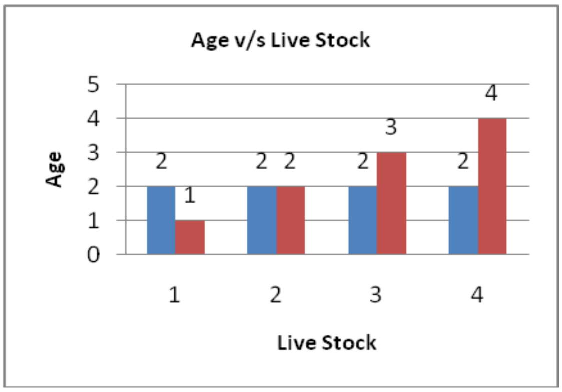

Fig 5:Age v/s Livestock (Threshold values as per fig.3)

Graph shown in fig.6 indicates the mutual relationship in between age and risk willingness attitude and depicts that 2% farmers of age group between 50 years to 75 years have more risk willingness than the farmers of age less than 50 years but it also shows that different level of risk willingness are available in the age level of 25 years to 50 years of farmers.

Age	Risk willingness	Support(%)
1	2	21
2	2	44
3	2	32
4	3	21

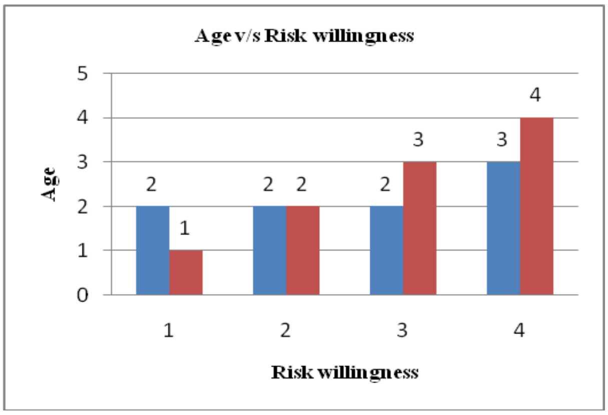

Fig 6: Age v/s Risk willingness (Threshold values as per fig 3)

Graph shown in fig 7 indicates the mutual relationship in between education and risk willingness and depicts that 2% farmers with education level of 10th class have more risk willingness than the farmers with education level of 8th class. Farmers with education level of 8th class shows different degree of risk willingness.

Education	Risk willingness	Support(%)
2	1	21
2	2	44
2	3	32
3	4	2

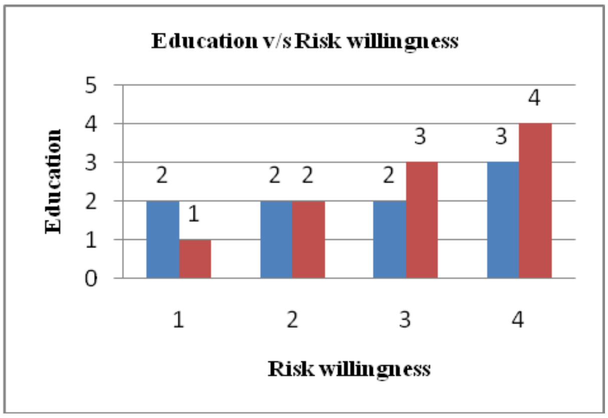

Fig 7:Education v/s Risk willingness (Threshold values as per fig.3)

Graph shown in fig.8 indicates the mutual relationship in between experience and live stock and shows that less experienced farmers possessed less live stock and more experienced farmer possessed different level of live stock. For instance, farmers having 10 to 20 years of experience possessed different level of live stock indicated by the threshold value of 1, 2 and 3.

Experience	Live stock	Support (%)
3	1	40
3	2	33
3	3	22
2	4	3

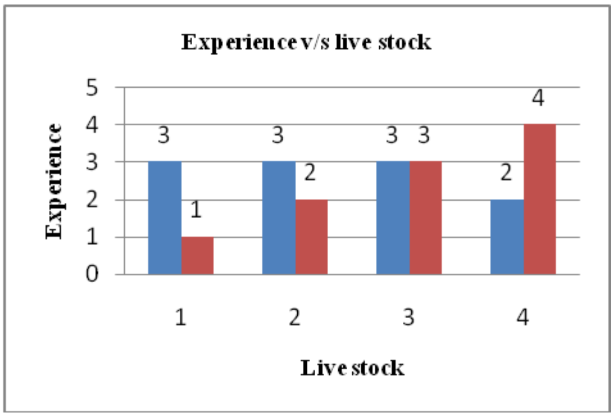

Fig 8: Experience v/s live stock (Threshold values as per fig 3)

Graph shown in fig.9 indicates the mutual relationship in between experience and extension-services and shows that inexperienced farmers (3%) used these services more in comparison to more experienced farmers.

Experience	Extension-services	Support(%)
3	1	16
3	2	30
3	3	50
2	4	3

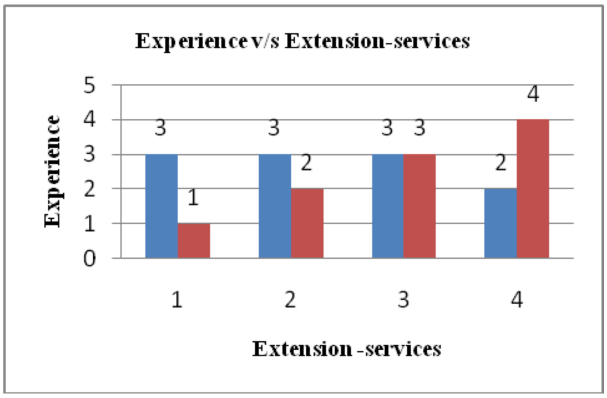

Fig 9: Experience v/s Extension-services (Threshold value as per fig 3)

Graph shown in fig.10 indicates the mutual relationship in between experience and risk willingness and conclude that experienced farmers are less risk taker.

Experience	Risk willingness	Support (%)
3	1	21
3	2	44
3	3	32
1	4	2

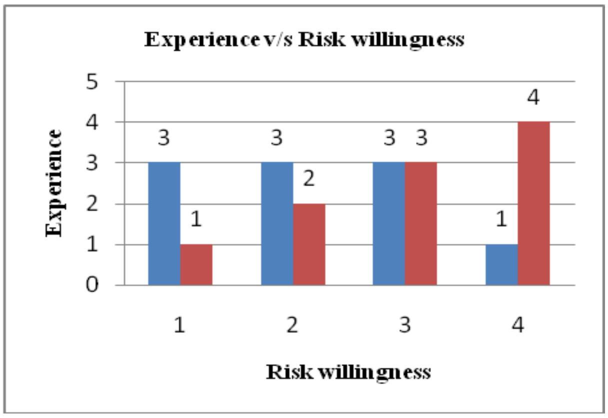

Fig 10: Experience v/s risk willingness (Threshold values as per fig.3)

Graph shown in figure 11 indicates the mutual relationship in between family dependents and live stocks and results that the farmer with more family dependents possessed more live stocks.

Dependents	Live stock	Support (%)
2	1	40
3	2	33
3	3	22
4	4	3

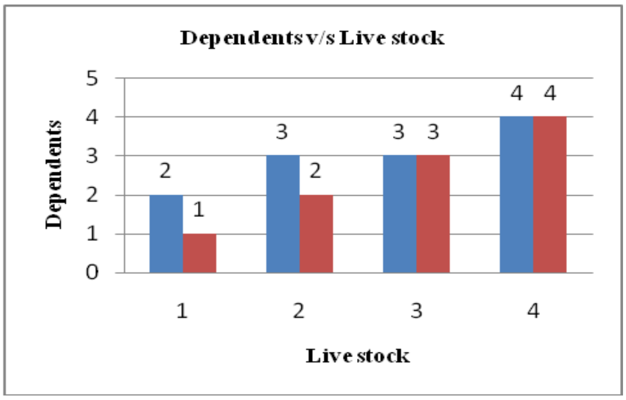

Fig 11: Dependents v/s live stock (Threshold value as per fig 3)

Graph of fig.12 displays the mutual relationship in between family dependents and extension-services and shows that the innovative attitude increases with the increment in family dependents.

Dependents	Innovativeness	Support (%)
2	1	17
3	2	35
2	3	47
0	4	0

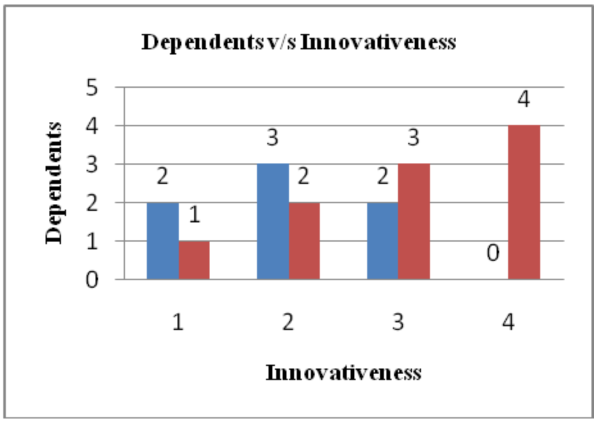

Fig 12: Dependents v/s Innovativeness (Threshold value as per fig.3)

Graph shown in fig.13 indicates the mutual relationship between live stock and land size and indicates that the farmers with bigger land size possesses more live stock than the farmers with smaller land size.

Live stock	Land size	Support(%)
1	1	22
1	2	30
2	3	15
2	4	32

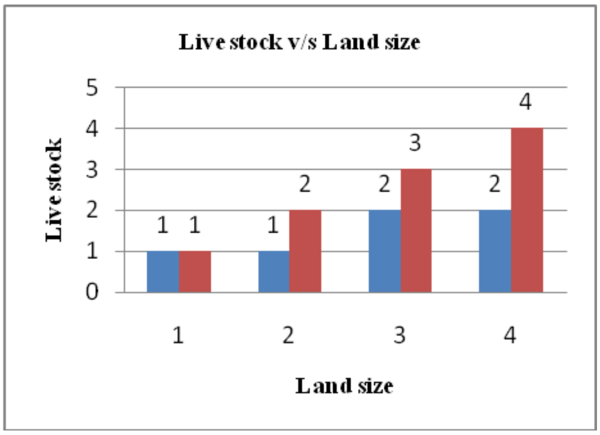

Fig. 13.Live stock v/s Land size (Threshold value as per fig.3)

Graph shown in figure 14 indicates the mutual relationship in between innovativeness and extension services and concludes that more innovative farmers use more extension-services.

Innovativeness	Extension-services	Support(%)
1	1	16
1	2	30
2	3	50
2	4	3

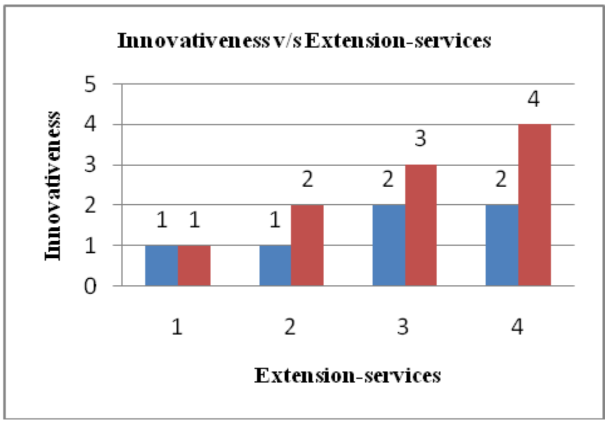

Fig 14: Innovativeness v/s Extension-services (Threshold value as per fig.3)

4. CONCLUSION AND FUTURE WORK

Data mining is a new discipline with a wide variety of techniques and their applications. There is a big gap in between the existing data mining applications and the need of industrialist, scientists and individuals. Data mining systems must be more users friendly, interpretative and explorative towards the handling of more complex data. In this paper, an algorithm to find the mutual relationship in IDIV affecting farmers' income is given and discussed. Results conclude that some variables are mutually dependent on some other variables for instance live stock and land size, with the increment in the land size, live stocks also increases. Similar relationships are found in family dependents and live stock, education and risk willingness and innovativeness and extension-services. In some IDIV negative relationships are found for instance less experience farmers use extension services more and more experiences farmers take less risk. The avenues of future work include developing more advanced pattern mining algorithm including mining multidimensional relationship among item sets, mining multiple level association rules and also mining mutual relationship with correlation among item sets.

5. REFERENCES

[1] 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.

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

[3] 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.

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

[5] Frayyad U. and Piatetsky-Shapiro G.Smyth P,"Knowledge discovery and data mining:Towards a unifying framework. In:Proceedings of the 2nd ACM international conference on knowledge discovery and data mining(KDD),Portland,OR,pp 82-88, 1996.

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

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

[8] 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.

[9] 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.

[10] 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.