Constraint based Data Mining Focusing Farming Case Study

Mamta
Shobhit University
Meerut

Shwetank Arya
Gurukul Kangri University
Hardwar

R. P. Agarwal
Shobhit University
Meerut

ABSTRACT

Data mining process may uncover thousands of patterns from a given data set; most of them may be unrelated to the users' interest. Also these rules occupy more memory space, take more time also require more efforts of the decision maker in analysis. To confine the search space users have the good sense of which direction of mining may lead to related or interested patterns they would like to find. Therefore, a good heuristic is to have the users specify such intuition or expectation as constraints to limit the search space. In this paper efforts are made to discover valuable patterns using the user input constraints.

General Terms

User specific constrained based data mining.

Keywords

Threshold value, Constraint Based data mining, IDIV.

1. INTRODUCTION

Data mining process surface thousands of rules from a given data set, most of which may be uninteresting or unrelated to the need of the user. Often, users have the good sense of which line of mining may lead to related or desired patterns. Therefore, a good heuristic is to have the users specify such intuition or expectation as constraints to confine the search space. This strategy is known as constraint-based data mining [1]. The constraints can include the following users' choices.

- Knowledge type constraints: These constraints specify the type of knowledge to be mined, such as characterization, discrimination, association and correlation analysis.
- Dimension/level constraints: These constraint specify the desired dimensions (or attributes) of the data, or levels of the concept hierarchies to be used in mining.
- Levels of concepts hierarchies: Concepts hierarchies are a popular form of background knowledge, which allow data to be mined at multiple levels of abstractions.
- Interestingness constraints: These constraint based on the threshold value of support or confidence. Rules whose support and confidence values are below from the user-specified thresholds are considered uninteresting.
- Meta rule constrained: These specify the form of rules to be mined. Meta rule may be based on the user's knowledge, understanding, experience or expectations. Generally, meta-rule forms a hypothesis regarding the relationships that the user is interested in confirming. The data mining system can then search for rules that match the given meta-rule.

Constraint based data mining allows users to describe the rules that they would like to focus, thereby making the data mining process more relevant and effective. Constraints can be implemented using high-level declarative data mining query language, user interface or query optimizer. In the proposed mechanism[Fig 1.] user interface is used to accept the users choices, To assess efficiency of the proposed mechanism the data set related to socio-economic conditions of farmers is used to mine multi dimensional association rules. Data set consists of threshold values [Table 3] of 9 socio-economic inter disciplined independent variables (IDIV) affecting farmers' income per annum [7-8]. Income is given in thousands.

2. Related Work

In a study collaborative filtering technique was used in a personalized recommendation model designed for web mining [2]. Researchers developed the recommender engine based on association rules for distributed environment that facilitate the system expansion and redistribution between hosts [3]. A top-down progressive deepening technique is developed for mining multiple level association rules [4]. A study is done to know how the filtering performance for a machine learned filter is affected when users explicitly modify the filtering profile [5]. Recommendation systems have become a popular tool for facilitating users in searching valuable information. Researches in this area focused on developing algorithms for efficiently producing recommendations to users. But effective explanation of these recommendations is also an important issue to increase the adoption and satisfaction level of users [9]. In this paper a constrained based mechanism is discussed which allow the user to input the constraint to discover the desired patterns.

3. Proposed Work

Constrained based Intelligent Data Mining Mechanism (IDMM) is proposed to help the users finding relevant and valuable information. The system consists of four modules: User, dialog management, inference engine and data repository [Fig 3.]. The model can be used in various applications such as e-commerce, education, farming applications etc. In the present study the model is assessed on the real world data set related to farming conditions. Dataset is collected through questionnaire from 324 farmers located in villages near to Meerut city. The Mechanism can be used to guide various users associated with farming such as farmers, NGOs and government organization personnel working for the growth of farmers and farming products. The aim of the proposed model is to find the most relevant information for the satisfaction of the user to improve the farmers' income and agricultural productivity.

3.1 Inter disciplined independent variable (IDIV):

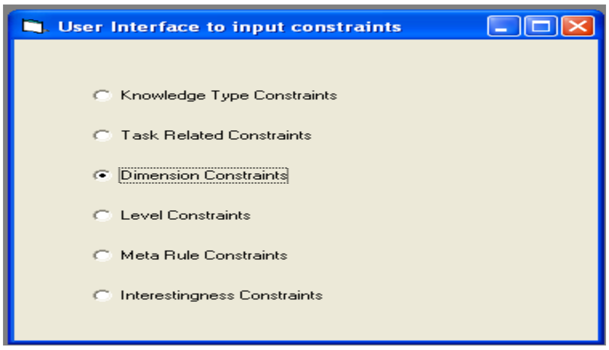
Present study introducing the new type of variables named Inter Disciplined Independent Variables (IDIV). IDIV are those variables which affect other variables with their presence. In real life applications such as medical, agriculture, education, sale purchase and many more, certain symptoms, behavior, performance and practices depends upon various inter related factors. For instance, in medical a specific disease occurred due to various factors. Different socio-economic conditions are responsible for the academic performance of the students [10]. Similarly, in the field of agriculture farming practices affected from various socio-economic conditions. Such distinguished factors not only affect the dependent activity but also affect other factors and also get affected from the presence of other factors. Such types of variables are named as inter disciplined independent variables. Another advantage of introducing IDIV is that these variables can be assigned threshold values which help to make algorithm more general and also help to analyze the results.

3.2 Constrained based Intelligent Data Mining

Constraint based mining process comprises following components:

3.2.1. Dialog Management

Dialogue management module facilitates the users to select the given constraint and support the user to decide how and what type of constraints should be selected to mine the relevant information. For instance, if farmer is using the system to gain the knowledge of innovative techniques, Mechanism display the list of constraints [Fig 1]. Option 3 displays the dimension level constraints. System further display the list of attributes to allow the farmer to selects the attributes [Fig 2.].



User Interface to input constraints

○ Knowledge Type Constraints

○ Task Related Constraints

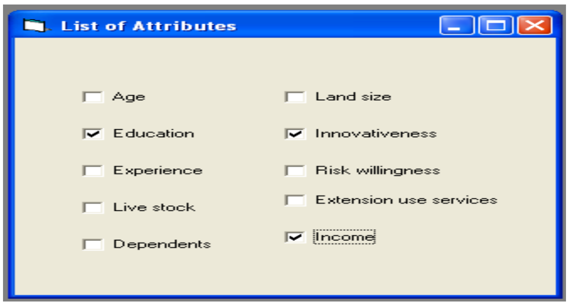
● Dimension Constraints

○ Level Constraints

○ Meta Rule Constraints

○ Interestingness Constraints

Fig 1: List of constraints



List of Attributes

☐ Age

☑ Education

☐ Experience

☐ Live stock

☐ Dependents

☐ Land size

☑ Innovativeness

☐ Risk willingness

☐ Extension use services

☑ Income

Fig 2.List of attributes

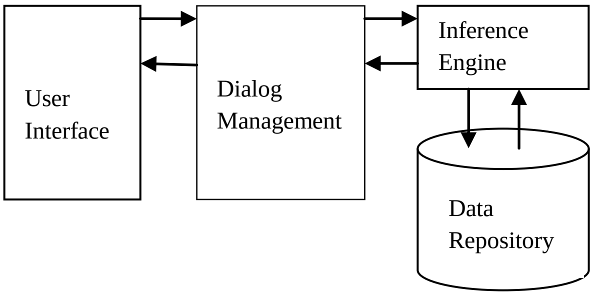


Fig 3. User specific constrained based data mining mechanism.

The dialog management component is responsible for making dialogs with the users presenting the queries to the users and processing the users' responses to propagate the further queries to the users and also to input the knowledge gained from the response to the next inference module. For instance if farmer responded association analysis, next question will be propagated to the farmer to select the attributes in which the user need the associations.

3.2.1User Module

User module provides the user friendly input interface and allow the user to interact with the system.

3.2.2 Inference Engine

The Inference engine is responsible for doing the mapping in between dialog management module and the attributes necessary to discover the desired patterns. Inference engine accept the output from the dialog management module which is free from all uncertainty and noise and do the mappings.

3.3.3. Data Repository

Data repository contains 9 IDIV and one dependent variable, income. These 9 IDIV are identified as important determinants by many researchers [7-8]. All 9 IDIV are assigned 4 threshold values 1 to 4 to provide the generality to the model. Age IDIV refers the age levels of farmers. Farmers below 25 years, threshold value 1 is assigned, Farmers below 50 years but more than and equal to 25 years, 2 threshold value is assigned. Farmers below 75 years but more than and equal to 50 years, threshold value 3 is assigned and farmers having the 75 years or more are assigned threshold value 4. Second IDIV is education. Farmers belonging to different educational standards, different threshold values are assigned. Farmers having the education less than 8th class, threshold value 1 is given, farmers having the education below 10th standard but more than and equal to 8th standard, threshold value 2 is assigned, farmers belonging to below 12th standard but more than and equal to 10th standard, threshold value 3 is assigned and farmers more than or equal to 12th standard, threshold value 4 is assigned. Third IDIV is experience. Farmers with experience below 10 years, threshold value 1 is given, below 20 years but more than and equal to 10 years, 2 is given, farmers below 30 years of experience but more than and equal to 20 years experience, 3 is assigned and farmers more than and equal to 30 years, threshold value 4 is assigned. Fourth IDIV is dependents. It refers to family dependent. Family dependent is also considered important determinant. It is assumed if families dependent are more; farmer may be more concerned towards the farming practices. Farmers having the family dependent below 3, threshold value 1 is assigned, more than and equal to 3 but less than 6, 2 threshold value is assigned, more than and equal to 6 but less than 8 dependent, threshold value 3 is assigned and farmers having the dependents more than or equal to 8, threshold value 4 is assigned. Live stock is fifth IDIV. It refers to the live stocks used for farming. Farmers having the live stock less than 3, threshold value 1 is assigned, Live stock more than or equal to 3 but less than 6, threshold value 2 is assigned, Live stock more than or equal to 6 but below 8, threshold value 3 is assigned and live stock more than or equal to 8, threshold value 4 is assigned. Land size is seventh IDIV. It is assumed that farmers having the bigger land size will have more live stock and more innovative attitude. Lands sizes vary in area and according to area threshold value are assigned from 1 to 4. The next IDIV is extension-services. These services are provided by the organizations or experts who initiate activities to help farmers in using new techniques and resources in farming. Some farmers never use these services, some occasionally use, some generally use and some are regular user of these services. Accordingly, 1 to 4 threshold values are assigned to extension services IDIV. The next IDIV is innovativeness. Innovativeness refers to apply new ideas and practices in farming. Some farmers always apply traditional methods, some use innovative methods, some other generally use new methods and resources and some other farmers are very innovative and always apply new ideas and methods. According to their innovativeness 1 to 4 threshold values are assigned to this IDIV. The last IDIV is risk willingness. Some farmers never take risk, some sometimes take risk, some generally take risk and some very often take risk. According to their risk willingness 1 to 4 threshold values are assigned. The complete list of IDIV and threshold values is given in table 3.

Example: User is asked to select the constraint from the displayed window [Fig 1]. If user selects the support 4% and confidence 15%, then the system display more valuable information in comparison to case when user search without giving percentage of support and confidence. System displayed information for 8 results when percentage of support and confidence is given[table 2] and 14 results when this information is absent [Table 1].

4. Result Evaluation

The mechanism is tested on the data set stored in farmers' repository. Various IDIV are existed in the repository. Two IDIV are education and extension-services, used to assess the performance of constrained based mining process using the algorithm [11].

5.1 Education and Extension-services and dependent variable farmers' income per annum

Table 1. [14 Results when no value for support and confidence is given, threshold values of Edu (Education) and ES (Extension-services) as per table 3]

Education	Extension-Services	Income (Th)	Support (%)	Confidence (%)
1	2	323	4	88
1	3	250	.006	11
2	1	281	4	17
2	2	256	8	39
2	3	239	9	41
2	4	300	.003	1
3	1	214	4	14
3	2	178	5	44
3	3	172	12	37
3	4	185	1	3
4	1	123	.92	2
4	2	180	19	50
4	3	148	16	43
4	4	292	1	4

Table 2 [8 Results when support is 4% and confidence 15%,Threshold values of Edu (Education)and ES (Extension-services) as per table 3]

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	23	180	19	50
4	3	148	16	43

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

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 &	3

	Dependents < 8	
	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 Willingness	No risk willingness	1
	Risk willingness	2
	More risk willingness	3
	Most risk willingness	4

Conclusion

The constrained based data mining technologies have generated new dimensions and opportunities of customization. Constraints limit the Search space and help computational statements to specify where to begin, How to calculate the path to descent and when to terminate the search. In this paper issues related to customization are discussed in the light of constrained based mining of data from large data set. New types of variables named inter disciplined independent variables are discussed and threshold values are assigned to these variables to make the results more valuable. The paper discussed user specified interactive explorative constraint data mining that is assessed on real world data set related to socio-economic conditions of farmers and resulted that constrained based data mining mechanism produce more valuable, concise and concrete results as compare to non constraint based data mining.

5. REFERENCES

[1] Kamber M. and Han J., ”Data Mining Concepts and Techniques”, Edition 2nd, 2010.

[2] Wand T. and Ren Y., ”Research on Personalized Recommendation Based on Web Usage Mining using Collaborative Filtering Technique”, WSEAS Transactions on Information Science and applications, 2009.

[3] Kazienlo P., “Mining Indirect Association Rule for Web Recommendation”, Int. J. Appl. Math. Computer Sci., vol 19, no.1, 2009.

[4] Han J., “Mining Multiple-Level Association Rules in Large Databases”, IEEE Transactions on Knowledge and data engineering, Vol 11, No. 5, 1999.

[5] Waern A., “User Involvement in Automatic Filtering: An Experimental study”, User Modeling and User-Adapted Interaction, 14:2004.

[6] Andonie R., Dean R. and Russo J.E., “Crossing the Rubicon for An Intelligent Advisor”, Proceedings of a workshop on the next stage of researches, San Diego, 2005.

[7] 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.

[8] 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.

[9] Andonie R., Russo J.E. and Dean R., ”Crossing the Rubicon for Intelligent Advisor”, Proceedings of a workshop on Beyond Personalization, San Diago, Jan 2005.

[10] Mamta, S. Arya and R.P. Agarwal, “Web Based Decision Making System for Assessing Socio-Economic Problems”, Proceedings of 2nd National Conference on Global Trends & Innovations in Computer Applications and Informatics, Shobhit University, Meerut, 2011.

[11] Mamta, S. Arya and R.P. Agarwal, “Identification of Multidimensional Relationship among item sets using association rules”, Journal of Information Science, May 2012