International Journal of Soft Computing 2 (2): 243-248, 2007
© Medwell Journals, 2007

Assessing the Interestingness of Discovered Knowledge Using a Hybrid Approach Based on Fuzzy Concepts

[1]R. Radha and [2]S.P. Rajagopalan
[1]Department of Computer Science, S.D.N.B. Vaishnav College For Women Chromepet, Chennai-44, India
[2]Mohamed Sathak, Group of Educational Instititiutions, Chennai-5, India

Abstract: A data mining technique usually generates a large amount of patterns and rules. However, most of these patterns are not interesting from a user's point of view. Beneficial and interesting rules should be selected among those generated rules. This selection process is what we may call a second level of data mining. To prevent the user from overwhelming with rules, techniques are needed to analyze and rank them based on their degree of interestingness. There are two aspects of rule interestingness, objective and subjective aspects. In this study, we are concentrating on both the subjective and objective measures of interestingness. A generic problem of finding the interesting ones among generated rules is addressed and a new mathematical measure for finding the interestingness is explained. We have used fuzzy linguistic terms for the attributes, so that the semantics of such rules are improved by introducing imprecise terms in both the antecedent and the consequent, as these terms are the most commonly used in human conversation and reasoning. The terms are modeled by means of fuzzy sets defined in the appropriate domains. However, the mining task is performed on the fuzzy data. These fuzzy association rules are more informative than rules relating precise values.

Key words: Fuzzy association rules, subjective rule interestingness, rule interestingness, fuzzy linguistic terms, Apriori MR, post-analysis of rules

INTRODUCTION

An earlier attempt addressing this rule of interestingness issue in (Knowledge Discovery in Databases) KDD systems has been presented in (Gregory and Christopher, 1994). A more general discussion is available with (Avi and Alexander, 1996), where Silberschatz and Tuzhilin propose unexpectedness and actionability as user-oriented measures of pattern interestingness, which has been further elaborated in (Balaji and Alexanr, 1998). A data-driven aspect, is based on statistics and structures of patterns, e.g., support, confidence, etc. On the other hand, subjective interestingness is user driven, it is based on user's belief in data, e.g., unexpectedness, novelty, actionability, etc. For instance, subjective approaches can be used as a kind of first filter to select potentially interesting rules, while objective approaches can then be used as a final filter to select truly interesting rules (Freitas, 1999). One should note that both aspects should be used to select interesting rules, since a data mining process mostly needs subjective evaluation of the domain.

There is an increasing interest in finding association rules among values of quantitative attributes in relational databases (Srikant and Agraual, 1996), as these kind of attributes are rather frequent. Quantitative attributes are those whose domain contain many precise values. Medical databases are used to store a big amount of quantitative attributes. But in common conversation and reasoning, humans employ rules relating imprecise terms rather than precise values. For instance, a physician will find more appropriate to describe his/her knowledge by means of rules like if fever is high and cough is moderate then disease is X than by using rules like if fever is 39.78°C and cough is 6 over 10 then disease is X. It seems clear that rules relating precise values are less informative and most of the time they seem strange to humans. Our goal is to find association rules with improved semantics (i.e., relating imprecise terms with clear semantic content) from a database containing precise values.

An inherent weakness of the objective measures is due to the fact that they do not consider the human background knowledge about the application domain. In this, domain knowledge is nothing but anything an expert knows about the particular application selected like prior expectations, intuitions and framed knowledge which he gained through his experience like a doctor gains experience in diagnosing. This study is explained as subjective interestingness. In this study an attempt is

Corresponding Author: R. Radha, Department of Computer Science, S.D.N.B. Vaishnav College For Women Chromepet, Chennai-44, India

made to build the gap between the qualitative and somewhat intuitive human knowledge and the quantitative results returned by mining algorithms based on fuzzy sets. The data obtained are converted to fuzzy values which in turn are expressed in terms of fuzzy linguistic terms so that rules like a man above 45 may have chance of blood pressure 160/100 can be expressed as a middle age man can have high blood pressure. Here middle age, high are considered to be fuzzy linguistic terms.

Association rules give strong rules for binary tuples. But most of the real world data are continuous numerical attributes which in turn when attempted to convert into binary value will lose the real meaning or information which the original has. For example, a student database consists of student marks, name, course etc. In an attempt to convert this database to binary, i.e., let us take student marks alone, if the mark field is converted to 1 for those who get above 40 marks and others as 0 then we lose the information that who is distinguished, first class, second class, average etc. Not only that we lose the degree of membership also

A system has been implemented to perform the post analysis of association rules generated by systems such as Apriori MR. Past research on inductive learning has mostly been focused on techniques for generating concepts or rules from datasets (Quinlan, 1992; Micharski, 1980; Clark and Nibleet, 1989). Limited research has been done on what happens after a set of rules has been induced. It is assumed that these rules will be used directly by an expert system or some human user to infer solutions for specific problems within a given domain. Having obtained a set of rules is not the end of the story, post-analysis of rules has to be done. The motivation for performing post-analysis of the rules comes from realizing the fact that using a learning technique on a dataset does not mean that the user knows nothing at all about the domain and the dataset. This is particularly true if the user is a human being. Typically, the human user does have some pre-conceived notions or knowledge about the learning domain. Hence, when a set of rules is generated from a dataset, naturally he/she would like to know the following: Do the generated rules represent what I know already? If not, which part of my previous knowledge is correct and which part is not? In what ways are the new rules different from my previous knowledge? Past research has assumed that it is the user's responsibility to analyze the rules to answer these questions. However, when the number of rules is large, it is very hard for the user to analyze them manually.

A window-based user interface has been implemented in VB to get the input from the user in terms of continuous attributes like age, fine etc as most of the data we get are numerical values. These data are converted to fuzzy linguistic terms with the help of fuzzy membership generation for each attribute. Some of the inputs like accident location i.e., whether they took place at bend, straight, a junction etc can be get as discrete values. This prepared database is given as input to a data mining algorithm for generating the association rules. This will generate a vast number of rules of which most of them are not interesting. The discovered rules are again fed to the database which are filtered with different measures of objective interestingness and are given as output to the user in the form of report. In this different measures like PS, RI, support, Confidence are discussed. Along with it the newly constructed measure SQSPR is calculated and the significance of this measure over the previous measure are discussed by presenting the calculated values in a tabular form.

For this purpose a case study has been carried out over a driver data base. In this study, rules are learned using the association rule generator AprioriMR association rule (MR researches 2004) .Different measures like the confidence , support , PS , IS and SQSPR values are calculated for the discovered rules and the results are displayed.

Proposed measures: While it is important to generate understandable rules, it is also important to the domain experts to have a complete picture of all the rules that exist in the database. This leads naturally to association rule mining. The problem with association rule is, however, that there are often too many of them, and they also contain a large amount of redundancy (Lia *et al.*, 1999) Past research in dealing with this problem can be described with the following approaches:

- Discover all the rules and select only the user interested rules to be stored in templates (Klemetinen *et al.*, 1994).
- Use constraints to constrain the mining process to generate only relevant rules (Srikant *et al.*, 1997).
- Find unexpected rules. This approach first asks the user to specify his/her existing knowledge about the domain.

Piatetsky-Shapiro proposed three principles for rule interestingness (RI), which can be stated as follows (Klemetinen *et al.*, 1994).

- RI = 0, if |A and B|=|A||B|/N (Here N is the number of tuples. Rule interestingness is 0 if rule antecedent and rule consequent are statistically independent.)
- RI monotonically increases with |A and B| when other parameters are fixed.
- RI monotonically decreases with |A| or |B| when other parameters are fixed.

Freitas has pointed out five additional factors related to the interestingness of a rule (Freitas, 1999). These are disjunct size, imbalance of the class distribution, attribute costs, misclassification costs and asymmetry in classification rules. These factors are claimed to be important when designing a domain-dependent rule interestingness measure.

Small disjunct problem, as it is addressed in (Freitas, 1999), is related to the tendency of most data mining algorithms to discover large disjuncts. This bias towards large disjuncts can eliminate very important and interesting knowledge if not treated properly. Especially, in domains where small disjuncts collectively match a large percentage of the tuples, this is an undesired situation. In order to overcome this problem, small and large disjuncts should be evaluated in different ways.

In some of the domains, the tuples belonging to one class are much more frequent than the tuples belonging to other classes. This is called imbalance of the class distribution. In such a study, the smaller the relative frequency of a minority class, the more difficult to discover rules predicting it. So, from the interestingness point of view, the rules predicting the minority class are more interesting.

Freitas has also mentioned the fact that most rule interestingness measures consider the rule antecedent as a whole. However, as he states, the interestingness of two rules having the same coarse-grained value may be different, depending on the attributes occurring in the rules antecedent. So, a good interestingness measure should consider attributes individually, according to their specific properties that may be domain dependent. Attribute costs are one of these properties. In some application domains such as medical diagnosis, different attributes might have very different costs. In such a case, a rule whose antecedent consists of less costly attributes are more interesting.

Misclassification cost is another issue, which should not be ignored when designing a good interestingness measure. Especially in domains where misclassifying a case is highly crucial, users of the domain are in need of a less risky classification system. In order to achieve such a system, misclassification costs of the produced rules should be reasonable. In other words, the smaller the misclassification cost of a rule, the more interesting it is.

The last factor that has been stated by Freitas is the asymmetry in classification rules. A rule interestingness measure is said to be symmetric with respect to the rule antecedent and the rule consequent. The reason for this is that we want to discover rules where the value of predicting attributes determine the value of the goal attribute. Besides all these, there are several other rule interestingness measures in the literature. Some of these measures are discussed in (Hilderman and Hamilton, 1999). Most of these measures depend on statistical factors such as correlation.

Pruning redundant rules: If the data mining technique used for extracting rules produces redundant rules, this redundancy should be eliminated before calculating interestingness of rules. Here we may basically say a rule is redundant if it satisfies one of the following two conditions:

- If there are two implications of the form A→C and A and B→C and both rules have similar confidence values, then the rule A and B →C is redundant.
- If there are two implications A→C and B→C, both have similar confidence values, then B→C is redundant if B is a subset of the conditions of A.

The first principle says that if the addition of one condition to the rule antecedent does not affect the confidence of a rule, then, addition of that condition is unnecessary. The second principle says that the subsets of a generated rule are redundant if they are of the similar confidence strength.

Tan and Kumar also argued that a good interestingness measure should take into account the support of the pattern (Tan and Kumar, 2000). They have showed that their proposed IS measure can be used in the region of low support, i.e., support of 0.3, whereas using RI measure in the region of high support is preferred. Hence, in order to make small disjuncts as interesting as large disjuncts, we may take IS measure as the basic measure for rules having coverage values in the range of [0,0.3] and RI measure for rules with coverage values [0.3,1]. Here are the formulations for two measures, respectively:

$$\mathrm{IS} = \frac{\sqrt{P(A,B)P(A,B)}}{P(A)P(B)}$$

$$\mathrm{RI} = P(A,B) - P(A)P(B)$$

A study and implementation: In this study our data set consists of 50 cases. Aprior MR association rule algorithm is used to generate rules. In this the confidence, support, conviction and surpringness values are calculated. Along with it the SQSPR value is calculated.

If the Apriori MR algorithm is used it generates nearly 624 rules. But when we use supervised association rules, the rules are filtered and in each category only few rules are generated. The rules are generated based on a defined target class. The insurance company may be interested in giving preference to those who pay the fine less, which depends on so many criteria like age, accidentlocation (bend , straight, junction), no of accidents done in the previous history etc. In addition of specifying minimum confidence and minimum support one more new threshold called SQSPR is used.
The measure is given as

$$\mathrm{SQSPR} = \sqrt{\frac{P(A\cap B)}{P(A)} - \frac{P(A\cap B)}{P(B)}}$$

Using this measure the rules can be further filtered using the above measure.

Suppose there are two items {A,B} where A->B has a support of 15% and a confidence of 60%. Because these values are high, a typical association rule algorithm probably would deduce this to be a valuable rule. However, if the probability to purchase item B is 70%, then we see that the probability of purchasing B has actually gone down presumably because A was purchased. Thus, there appears to be a negative correlation between buying A and buying B. The correlation can be expressed as

$$\text{Correlation (A->B)} = \frac{P(A,B)}{P(A)P(B)}$$

which in this case is: Because this correlation value is

$$\frac{0.15}{0.25\times 0.7} = 0.857$$

lower than 1, it indicates a negative correlation between A and B.

But in SQSPR measure both the antecedent as well as consequent are given equal weightage and are evaluated.

- SQSPR = 0, If $\frac{P(A\cap B)}{P(A)} = \frac{P(A\cap B)}{P(B)}$

This shows that both antecedent and consequent are having equal contribution in framing the rule and are independent of each other and rules of these types can be eliminated.

- SQSPR increases if $\frac{P(A\cap B)}{P(A)} > \frac{P(A\cap B)}{P(B)}$

so the antecedent plays a major role in deciding the consequent

- SQSPR increases if $\frac{P(A\cap B)}{P(A)} < \frac{P(A\cap B)}{P(B)}$

This shows the negative correlation

These measure shows the strength and the interestingness of the discovered rule.

Data associations will be described by fuzzy rules, which extend the representational capabilities of classical association rules, facilitate the construction and interpretation of rules in natural linguistic terms, and avoid unnatural boundaries in the partitioning of the attribute domains.

There are two standard methods for extending crisp measures to fuzzy sets and fuzzy rules. The first, and perhaps simplest, is to directly replace the operators in the crisp measure with appropriate fuzzy counterparts. An alternative approach is to represent a fuzzy rule as a set of crisp rules. The confidence and support for the fuzzy rule are then obtained by applying standard techniques to the associated set of crisp rules and aggregating the results.

In this example out of 15 variables 7 are considered to be important as the deciding factors. This can be decided with the help of domain knowledge. This involves the first part of this paper. In the Apriori MR rule by specifying minimum confidence as 60% and minimum support as 10% about 620 rules are generated. This will be overwhelming. This discovered rules are again filtered by finding the SQSPR and specifying a minimum threshold and ranking them so that the user is presented with only 22 rules which will be easy to go through. The filtered rules are found to be interesting.

Experimental results

No.	Antécédent	Conséquent	PS	RI	SQSPR
125	"mstatus=married"-"farrests=zero"-"fuzzyage=adult"	"emp-uemp=employed"	0.0456	0.43994134	0.89802651
128	"sex=male"-"mstatus=married"-"finefuzzy=verylow"	"emp-uemp=employed"	0.0456	0.43994134	0.89802651
116	"emp-uemp=employed"-"finefuzzy=high"	"sex=male"	0.0456	0.43994134	0.89802651
121	"sex=male"-"farrests=zero"-"fuzzyage=adult"	"emp-uemp=employed"	0.0456	0.43994134	0.89802651
135	"emp-uemp=unemployed"-"fuzzyage=senior"	"faccidenloc=bend"	0.0432	0.4330127	0.901387819
94	"accidenloc=bend"-"sex=male"-"finefuzzy=verylow"	"mstatus=married"	0.044	0.42257712	0.906326967
95	"faccidenloc=bend"-"emp-uemp=employed"-"finefuzzy=verylow"	"mstatus=married"	0.044	0.42257712	0.906326967
96	"emp-uemp=employed"-"faccidenloc=straight"-"finefuzzy=medium"	"mstatus=married"	0.044	0.42257712	0.906326967
117	"mstatus=married"-"faccidenloc=straight"	"emp-uemp=employed"	0.038	0.40160966	0.91581094
110	"fuzzyage=adult"-"finefuzzy=low"	"emp-uemp=employed"	0.038	0.40160966	0.91581094
114	"faccidenloc=bend"-"sex=male"-"finefuzzy=verylow"	"emp-uemp=employed"	0.038	0.40160966	0.91581094
118	"sex=male"-"farrests=zero"-"fuzzyage=senior"	"emp-uemp=employed"	0.038	0.40160966	0.91581094
120	"farrests=zero"-"fuzzyage=adult"-"finefuzzy=low"	"emp-uemp=employed"	0.038	0.40160966	0.91581094
123	"sex=male"-"fuzzyage=senior"-"finefuzzy=verylow"	"emp-uemp=employed"	0.038	0.40160966	0.91581094
130	"faccidenloc=bend"-"farrests=zero"-"fuzzyage=adult"	"emp-uemp=employed"	0.038	0.40160966	0.91581094
106	"faccidenloc=bend"-"emp-uemp=employed"-"finefuzzy=verylow"	"sex=male"	0.038	0.40160966	0.91581094
107	"mstatus=single"-"finefuzzy=high"	"sex=male"	0.038	0.40160966	0.91581094
108	"farrests=three"	"sex=male"	0.038	0.40160966	0.91581094
109	"finefuzzy=high"-"farrests=two"	"sex=male"	0.038	0.40160966	0.91581094
126	"emp-uemp=employed"-"farrests=zero"-"fuzzyage=senior"	"sex=male"	0.038	0.40160966	0.91581094
113	"faccidenloc=straight"-"fuzzyage=senior"	"emp-uemp=employed"	0.038	0.40160966	0.91581094
136	"sex=male"-"mstatus=married"-"emp-uemp=unemployed"	"faccidenloc=bend"	0.036	0.3952847	0.918558654

The high ordered rule 125 says that if a person is married and he was not arrested so for and if he is employed then the insurance can be sanctioned to him without any risk.like wise other rules can be interpreted. The SQSPR gives the strength of the rules considering both the probability of antecedent as well as consequent of the rule.

No.	Antécédent	Conséquent	Confi	SQSPR
	"farrests=zero"-	"mstatus=single"	-	0.452267016
15	"fuzzyage=teenage"	"sex=female"	0.75	
	"emp-uemp=employed"-			
14	"finefuzzy=verylow"	"mstatus=married"	-	0.452267016
	"sex=male"-	"finefuzzy=medium"		0.5
	"farrests=one"	"farrests=zero"	0.75	
23			0.75	
218	"sex=male"-" mstatus=married"	"emp -uemp=employed"	0.75	0.515876954

In the above SQSPR the low value indicates that these are less interesting rules so the insurance company should not take risk in sanctioning insurance to these people. The rule 218 says that the person who is a male and married and even if he is employed he should not be given insurance shows some unexpectedness in this which again a rule evaluationary measure.

CONCLUSION

In this study a new objective measure say SQSPR is used and it proves to show the strength of the rule strongly. In this, based on the domain knowledge the data is selected by the user and then are mined to form the association rules using Apriori MR and then the derived rules are filtered using the SQSPR. In this, while getting the input from the user the some of the data are converted to fuzzy linguistic terms using fuzzy membership generation. Most of the work done is in the post analysis of the rules generated. A hybrid attempt of involving both subjective and objective measures of interestingness is applied to filter the novel, strong and interesting patterns.

REFERENCES

Gregory Piatesky-Shapiro and Christopher J. Matheus, 1994. The interestingness of deviations. In proceedings of KDD-94: AAAI-94 knowledge discovery in databases workshop, AAAI Press, pp: 25-36.

Avi Silberschatz and Alexander Tuzhilin, 1996. What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Eng., 8: 970-974.

Balaji Padmanabhan and Alexander Tuzhilin, 1998. A belief-driven method for discovering unexpected patterns. In proceedings of the international conference on knowledge discovery and data mining KDD, pp: 94-100.

Freitas, A.A., 1999. On rule interestingness measures. Knowledge-Based Sys., pp: 309-315.

Liu, B., W. Hsu and Y. Ma, 1999. Pruning and sumarizing the discovered associations. Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, KDD-99, San Diego, USA, pp: 125-134.

Klemetinen, M., H. Mannila, P. Ronkainen, H. Toivonen and A.I. Verkamo, 1994. Finding interesting rules from large sets of discovered association rules, CIKM.

Srikant, R., Q. Vu and R. Agrawal, 1997.Mining Association Rules with Item Constraints, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, KDD-97, Newport Beach, USA, pp: 67-73.

Srikant, R. and R. Agrawal, 1996. Mining quantitative Association Rules in Large Relational Tables. In: Proceedings of the 1996 ACM SIGMOD International Conference. Management Data, pp: 1-12.

Hilderman, R.J. and H.J. Hamilton, 1999. Knowledge discovery and interestingness measures: A survey. Technical Report CS 99-04, Department of Computer Science, University of Regina.

Quinlan, J. R., 1992. C4.5: Program for machine learning. Morgan Kaufmann.

Michalski, R., 1980. Pattern recognition as rule-guided induction inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2: 349-361.

Clark, P. and T. Niblett, 1989. The CN2 induction algorithm. Machine Learning, 3: 261-284.

Tan, P. and V. Kumar, 2000. Interestingness measures for association patterns: A Perspective. Technical Report # TR00-036, Department of Computer Science, University of Minnesota.