April 2014

FUZZY SET APPROACHES TO DATA MINING OF ASSOCIATION RULE

SHIVANGI SENGAR

IITM, Gwalior, shivangi26sengar@gmail.com

SAKSHI PRIYA

IITM, Gwalior, itsmesakshi.priya@gmail.com

URVASHI KHATUJA

IITM, Gwalior, urvashi.khatuja26@gmail.com

Follow this and additional works at: <https://www.interscience.in/ijcsi>

Part of the [Computer Engineering Commons](#), [Information Security Commons](#), and the [Systems and Communications Commons](#)

Recommended Citation

SENGAR, SHIVANGI; PRIYA, SAKSHI; and KHATUJA, URVASHI (2014) "FUZZY SET APPROACHES TO DATA MINING OF ASSOCIATION RULE," *International Journal of Computer Science and Informatics*: Vol. 3 : Iss. 4 , Article 3.

DOI: 10.47893/IJCSI.2014.1151

Available at: <https://www.interscience.in/ijcsi/vol3/iss4/3>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer Science and Informatics by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

FUZZY SET APPROACHES TO DATA MINING OF ASSOCIATION RULE

SHIVANGI SENGAR¹, SAKSHIPRIYA² & URVASHI KHATUJA³

^{1,2&3}IITM, Gwalior

Email:shivangi26sengar@gmail.com , itsmesakshi.priya@gmail.com, urvashi.khatuja26@gmail.com

Abstract:- Data mining on large databases has been a major concern in research community due to the difficulty of analyzing huge volume of data. This paper focuses on the large set area i.e. on fuzzy sets and knowledge discovery of data. Association rules* provide information in accessing significant correlations in large databases. We have combined an extended techniques developed in both fuzzy data mining and knowledge discovery model in order to deal with the uncertainty found in typical data.

INTRODUCTION

DATA MINING

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining* tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

FUZZY INFORMATION

Fuzzy sets are sets whose elements have degrees of membership. Fuzzy sets were introduced simultaneously by Lotfi A. Zadeh and Dieter Klaua in 1965 as an extension of the classical notion of set. In classical set theory, the membership of elements in a set is assessed in binary terms according to a bivalent condition — an element either belongs or does not belong to the set. By contrast, fuzzy set theory permits the gradual assessment of the membership of elements in a set; this is described with the aid of a membership function valued in the real unit interval $[0, 1]$. Fuzzy sets generalize classical sets, since the indicator functions of classical sets are special cases of the membership functions of fuzzy sets, if the latter only take values 0 or 1. In fuzzy set theory, classical bivalent sets are usually called *crisp* sets. The fuzzy set theory can be used in a wide range of domains in which information is incomplete or imprecise, such as bioinformatics

A fuzzy set is a pair (U, m) where U is a set and $m: U \rightarrow [0, 1]$.

For each $x \in U$, the value $m(x)$ is called the **grade** of membership of x in (U, m) . For a finite set $U = \{x_1, \dots, x_n\}$, the fuzzy set (U, m) is often denoted by $\{m(x_1)/x_1, \dots, m(x_n)/x_n\}$. Let $x \in U$. Then x is called not included in the fuzzy set (U, m) if $m(x) = 0$, x is called fully included if $m(x) = 1$, and x is called a **fuzzy member** if $0 < m(x) < 1$. The set $\{x \in U \mid m(x) > 0\}$ is called the **support** of (U, m) and the set $\{x \in U \mid m(x) = 1\}$ is called its **kernel**. The function m is called the **membership function** of the fuzzy set (U, m) .

FUZZY LOGIC

Fuzzy logic is a form of many-valued logic or probabilistic logic it deals with reasoning that is approximate rather than fixed and exact. In contrast with traditional logic theory, where binary sets have two-valued logic, true or false, fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions.

UNCERTAINTY OF DATA MINING

Data is often associated with uncertainty because of measurement inaccuracy, sampling discrepancy, outdated data sources, or other errors. This is especially true for applications that require interaction with the physical world, such as location-based services and sensor monitoring. For example, in the scenario of moving objects (such as vehicles or

people), it is impossible for the database to track the exact locations of all objects at all time instants. Therefore, the location of each object is associated with uncertainty between update. These various sources of uncertainty have to be considered in order to produce accurate query and mining results. In recent years, there has been much research on the management of uncertain data in databases, such as the representation of uncertainty in databases and querying data with uncertainty. However, little research work has addressed the issue of mining uncertain data. We note that with uncertainty, data values are no longer atomic. To apply traditional data mining techniques, uncertain data has to be *summarized* into atomic values. Taking moving-object applications as an example again, the location of an object can be summarized either by its last recorded location, or by an expected location (if the probability distribution of an object's location is taken into account). Unfortunately, discrepancy in the summarized recorded values and the actual values could seriously affect the quality of the mining results. Figure 1 illustrates this problem when a clustering algorithm is applied to moving objects with location uncertainty. Figure 1(a) shows the actual locations of a set of objects, and Figure 1(b) shows the recorded location of these objects, which are already outdated. The clusters obtained from these outdated values could be significantly different from those obtained as if the actual locations were available (Figure 1(b)). If we solely rely on the recorded values, many objects could possibly be put into wrong clusters. Even worse, each member of a cluster would change the cluster centroids, thus resulting in more errors.

Figure 1. (a) The real-world data are partitioned into three clusters (a, b, c). (b) The recorded locations of some objects (shaded) are not the same as their true location, thus creating clusters a', b', c' and c''. Note that a' has one fewer object than a, and b' has one more object than b. Also, c is mistakenly split into c' and c''. (c) Line uncertainty is considered to produce clusters a', b' and c. The clustering result is closer to that of (a) than (b).

FUZZY DATA MINING

Fuzzy set has been in pattern recognition, especially fuzzy clustering algorithms (Bezdek 1974*). Hence much of the effort in fuzzy data mining has been by the use of fuzzy clustering and fuzzy set approaches

in neural network and genetic algorithm (Hirota and Pedrycz 1999*). In fuzzy set theory an important consideration is the treatment of data from a linguistic view point from this has developed an approach that uses linguistically quantified propositions to summarize the content of a data base by providing a general characterization of the analyzed data (Yager 1991 Kacprzyk and Zadrozny 2000*). A common organization of data for data mining is the multidimensional data queue in data warehouses.

Fuzzy data mining for generating association rules has been considered by a number of researchers. There are approaches using the SETM* (set-oriented mining) algorithm (Shu et al 2001) and other techniques (Bose and Pivert 2001*) but most have been based on important Apriori algorithm. Extensions have included fuzzy sets approaches to quantitative data, hierarchies or taxonomies, weighted rules and interestingness measures. For our work, our main focus is on Apriori algorithm motivated by some of the above development.

ASSOCIATION RULES IN DATA MINING

Association rules* are if/then statements that help uncover relationships between apparently unrelated data in a relational database or other information repository.

An example of an association rule :

"If a customer buys a dozen eggs, he is 80% likely to also purchase milk."

Association rules are created by analyzing data for frequent if/then patterns and using the criteria *support* and *confidence* to identify the most important relationships. *Support* is an indication of how frequently the items appear in the database. *Confidence* indicates the number of times the if/then statements have been found to be true. In data mining association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

{onions, potatoes} $\Rightarrow$ {burger}

For example, the rule found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placement. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics. As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

Example database with 4 items and 5 transactions

transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Following the original definition by Agrawal et al the problem of association rule mining is defined as: Let $I = \{i_1, i_2, \dots, i_n\}$ be a set of n binary attributes called *items*. Let

$D = \{t_1, t_2, \dots, t_m\}$ be a set of transactions called the *database*. Each transaction in D has a unique transaction ID and contains a subset of the items in I . A *rule* is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short *itemsets*) X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

To illustrate the concepts, we use a small example from the supermarket domain. The set of items is $I = \{\text{milk, bread, butter, beer}\}$ and a small database containing the items (1 codes presence and 0 absence of an item in a transaction) is shown in the table to the right. An example rule for the supermarket could be $\{\text{butter, bread}\} \Rightarrow \{\text{milk}\}$ meaning that if butter and bread are bought, customers also buy milk.

Useful Concepts

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence.

- The *support* $\text{supp}(X)$ of an itemset X is defined as the proportion of transactions in the data set which contain the itemset. In the example database, the itemset $\{\text{milk, bread, butter}\}$ has a support of $1/5 = 0.2$ since it occurs in 20% of all transactions (1 out of 5 transactions).
- The *confidence* of a rule is defined $\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y) / \text{supp}(X)$. For example, the rule $\{\text{milk, bread}\} \Rightarrow \{\text{butter}\}$ has a confidence of $0.2/0.4 = 0.5$ in the database, which means that for 50% of the transactions containing milk and bread the rule is correct (50% of the times a customer buys milk and bread, butter is bought as well). Be careful when reading the

expression: here $\text{supp}(X \cup Y)$ means "support for occurrences of transactions where X and Y both appear", not "support for occurrences of transactions where either X or Y appears", the latter interpretation arising because set union is equivalent to logical disjunction. The argument of $\text{supp}()$ is a set of preconditions, and thus becomes more restrictive as it grows (instead of more inclusive).

- Confidence can be interpreted as an estimate of the probability $P(Y|X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS
- The *lift* of a rule is defined as $\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$

or the ratio of the observed support to that expected if X and Y were independent. The rule

$$\{\text{milk, bread}\} \Rightarrow \{\text{butter}\} \text{ has a lift of } \frac{0.2}{0.4 \times 0.4} = 1.25$$

- The *conviction* of a rule is defined as $\text{conv}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)}$

$$\{\text{milk, bread}\} \Rightarrow \{\text{butter}\} \text{ has a conviction of } \frac{1 - 0.4}{1 - 0.5} = 1.2$$

, and can be interpreted as the ratio of the expected frequency that X occurs without Y , if X and Y were independent divided by the observed frequency of incorrect predictions. In this example, the conviction value of 1.2 shows that the rule $\{\text{milk, bread}\} \Rightarrow \{\text{butter}\}$ would be incorrect 20% more often (1.2 times as often) if the association between X and Y was purely random chance.

Process

Frequent itemset lattice, where the color of the box indicates how many transactions contain the combination of items. Note that lower levels of the lattice can contain at most the minimum number of their parents' items; e.g. {ac} can have only at most $\min(a, c)$ items. This is called the *downward-closure property*.^[2]

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually divide up into two separate steps:

1. First, minimum support is applied to find all *frequent itemsets* in a database.
2. Second, these frequent itemsets and the minimum confidence constraint are used to form rules.

While the second step is straightforward, the first step needs more attention.

Finding all frequent itemsets in a database is difficult since it involves searching all possible itemsets (item combinations). The set of possible itemsets is the power set over I and has size $2^n - 1$ (excluding the empty set which is not a valid itemset). Although the size of the powerset grows exponentially in the number of items n in I , efficient search is possible using the *downward-closure property* of support also called *anti-monotonicity* which guarantees that for a frequent itemset, all its subsets are also frequent and thus for an infrequent itemset, all its supersets must also be infrequent. Exploiting this property, efficient algorithms (e.g., Apriori and Eclat) can find all frequent itemsets.

CONCLUSION

We have presented an approach to discovery of association rules for fuzzy spatial data where we are interested in correlations of spatial data. We have

combined and extended techniques developed in both spatial and fuzzy data mining in order to deal with uncertainty found in typical spatial data.

Some of our future work include hierarchies, weight and interestingness measures. Type II data of spatial relationship have cases that involve hierarchies. for example, the NEAR predicate could be organized as a hierarchy of relationships such as contains, intersects etc. Each of these might have a different strength in the hierarchy as well as being defined by fuzzymembership functions. The combination of these values is complex and must be worked out in the knowledge discovery context.

ACKNOWLEDGEMENT

we would like to thank our guide MR. PRADEEP YADAV, MISS NEHA JAIN, MISS NEHA SAXENA for their supervision, support and motivation throughout the work duration.

REFERENCES

- [1]. Bezdek J 1974 cluster validity with FUZZY SETS. Journal of cybernetics 3 : 5872
- [2]. Agarwal R, Imielinski T, and Swami A 1993 MINING ASSOCIATION RULES between set of items in large databases
- [3]. Bose P and Pivert O 2001 on some FUZZY EXTENSIONS of association rules
- [4]. Chawla S, Shekhar S, Wu W and Ozasmi U 2000 a framework for SPATIAL DATA MINING
- [5]. George R, Buckles, Petri F and Yazici A 1992 UNCERTAINTY MODELLING in object oriented geographical system
- [6]. Han J and Kamber M 2000 DATA MINING concepts and techniques
- [7]. Kacprzyk J and Zadrozny S 2000 on combining intelligent queering and data mining using FUZZY LOGIC concepts