Introducing Concept of Fuzzy Support Matrix for Interestingness Measures

Swati R.Ramdasi

Department of Computer Science, Pune Vidyarthi Griha's College of Science and Commerce, Pune 411009 Maharasthra, India * *jsp15@rediffmail.com*

Fuzzy association rules with its linguistic annotations and human interpretable form, has provided a convenient extension of association concepts to quantified attributes. The applicability is extended by combining extraction of both positive and negative association rules. Interestingness measures are used to filter out the useful and correct set of actionable association rules from the larger set of rules mined by association rule mining algorithms. Many measures such as Support, Confidence, Conviction and Certainty Factor, with their own area of applicability and statistical significance are popular. The wide range of measures is usually based on frequency counts or probability of occurrence of certain attribute patterns. Binary attributes uses a 2×2 contingency table as the basis for defining different measures. This paper presents concept of fuzzy support matrix using fuzzy partitions, as a natural extension of contingency table for the different interestingness measures. Those can be defined in a uniform and consistent manner. It uses the existing interestingness measures defined in new form using fuzzy support and illustrate these concepts using known data sets. This paper represent active research directions aimed at advancing the capabilities, applicability, and efficiency of fuzzy association rule mining in handling modern data challenges across various domains.

Keywords: Interestingness measures; Association Rules mining; Fuzzy sets.

Introduction

Association rule mining as introduced by Agarwal ([Agrawal et al. (1993)]) for Mar- ket Basket analysis was used to extract dependencies between binary attributes in super market data giving rise to association rules of the form ”If A is purchasedso is B”. These kinds of associations such as ”If personal computer is purchasedso is the printer”, were used to facilitate planning of marketing strategies such as product placement, cross selling, promotional pricing etc. Fuzzy association rules replaced interval rangesby linguistic variables and crisp intervals by fuzzy partitions. The fuzzy association rule in human interpretable form has rich applicability.

The Apriori algorithm is the basic algorithm and there are many variants and im- provements of the same ([Savasere et al. (1995)], [Brin et al. (1997)], [Hilderman and Hamilton (1999)]) that are mainly used for mining positive association rulesof the form $A \rightarrow B$. Negative association rules can be in one of following forms: $A \rightarrow \sim B$, $\sim A \rightarrow B$ and $\sim A \rightarrow \sim B$. There are several applications of negative association rules from drug discovery to error detection. The advantage of fuzzy partitioning is that the linguistic labels encompass both the positive and negative association rules into linguistic annotations of {Low, Medium, High} where ’Low’ represents absence or negligible presence while ’High’ indicates substantial presence giving rules which are more

closer to human interpretation such as ”If Tea is pur-chased in Low quantity then Coffee is purchased in High quantity”.
Association rule mining algorithms generate a large set of rules from which one needto filter out non-trivial, useful, interesting and actionable rules by assessing them using certain measures. Normally Support and Confidence measures are essentially used measures. The support measure gives statistical significance of a rule, where as confidence measure quantifies the strength of the rule. Subsequently, a large set of interestingness measures were added to the association rule mining landscape with very good properties and applicability ([Kamber and Shinghal (1996)], [Brin et al. (1997)], [Hilderman and Hamilton (1999)], [Bayardo Jr and Agrawal (1999)], [Tanet al. (2004)], [Geng and Hamilton (2006)]). These measures are usually defined for binary attributes using frequency counts tabulated in a 2 × 2 contingency table or using probability values again computed using frequency counts.
Fuzzy partitioning of quantified or ordered categorical attributes gives rise to ex- plosion in number of fuzzy association rules generated and researchers ([Ramdasi and Shirwaikar(2016)], [Rusnok and Burda (2017)]) have used extention of inter- estingness measures to limit fuzzy association rules. However applicability demands measurement of association between original atrributes and not between partitioned attributes.
As data streams become more prevalent, there is a need for incremental and online fuzzy association rule mining algorithms that can adapt to changing data dynamics in real-time or near real-time scenarios. Researchers continue to innovate to address the complexities and opportunities presented by fuzzy logic in discovering valuable insights from uncertain and imprecise data relationships. Interestingness measures help in filtering and selecting the most relevant and valuable fuzzy association rules from the potentially vast space of mined rules, ensuring that the discovered patterns contribute meaningfully to decision-making and knowledge discovery processes.

Recent developments and research trends in fuzzy association rule mining includes:

1. **Hybrid Approaches**: Researchers have been exploring hybrid approaches that combine fuzzy logic with other machine learning techniques such as neural networks, genetic algorithms, or swarm intelligence. These hybrids aim to improve the accuracy and efficiency of fuzzy association rule mining, especially in handling large-scale and complex datasets.
2. **Handling Big Data**: With the proliferation of big data, there is a growing emphasis on developing scalable algorithms and frameworks for fuzzy association rule mining. Researchers are working on methods to efficiently process and mine fuzzy patterns from massive datasets, considering both computational efficiency and memory management.
3. **Multi-level and Hierarchical Fuzzy Association Rules**: There is ongoing research into mining association rules at multiple levels of granularity or hierarchy, where fuzzy sets and fuzzy logic play a crucial role in capturing relationships between items at different abstraction levels. This approach is useful in various domains, including decision support systems and complex pattern recognition.
4. **Fuzzy Temporal Association Rules**: Temporal aspects are essential in many applications, such as analyzing time-dependent patterns in customer behavior or healthcare data. Recent research has focused on developing fuzzy temporal association rule mining techniques to handle fuzzy temporal intervals and uncertainty in temporal relationships.
5. **Applications in Healthcare and Bioinformatics**: Fuzzy association rule mining continues to find applications in healthcare informatics and bioinformatics. Researchers are exploring its potential in analyzing medical records, predicting diseases based on fuzzy patterns of symptoms, and understanding complex interactions in biological systems.
6. **Interpretability and Explainability**: Enhancing the interpretability and explainability of fuzzy association rules remains a significant area of research. Methods are being developed to visualize and present fuzzy rules in a meaningful way to domain experts, ensuring that the mined patterns are actionable and useful in decision-making processes.

Data mining involves the process of discovering patterns, correlations, anomalies, and trends within large datasets. It is a crucial part of extracting meaningful insights and knowledge from raw data. Association rule mining specifically focuses on identifying interesting relationships or associations between items in transactional databases or other types of data repositories. The primary motivation includes: **Market Basket Analysis**: Understanding which products are frequently purchased together to optimize product placement and promotions.

Fuzzy association rule mining extends traditional association rule mining by allowing for the representation of uncertain or imprecise relationships between items in datasets. Here's a survey of fuzzy association rule mining, covering its concepts, methodologies, applications, and challenges:

Fuzzy association rule mining addresses scenarios where relationships between items are not strictly binary (present or absent) but rather have degrees of membership or uncertainty. It integrates concepts from fuzzy logic to handle imprecise data effectively.

Background and Related work

A fuzzy set is identified by generalized characteristic function known as member- ship function [Zadeh (1965)]. The fuzzy partitioning for association rules has been discussed in more details by [Dubois et al. (2003)], [Dubois et al. (2006)]. The mem- bership function specifies the degree of membership μ in the fuzzy set. Given an attribute set A and linguistic label set L, the membership function μ is a mapping from A{L}→[0, 1]. There are several alternatives for membership functions whichcan be conveniently defined using mathematical formula for the complete attribute set and linguistic labels. A triangular membership function is specified using three parameters a, b and c and using min and max function as follows:

triangular(x: a, b, c) = $\max(\min(\frac{(x-a)}{(b-a)}, 1, \frac{(c-x)}{(c-b)}),0)$

Where x is attribute value and l is linguistic label. Only single value has full that is
1 membership and membership value goes on increasing from a to b and decreasesfrom b to c tending towards 0.
A trapezoidal membership has four parameters as described,

trapezoidal (x: a, b, c, d) = $\max(\min(\frac{(x-a)}{(b-a)}, 1, \frac{(d-x)}{(d-c)}),0)$

A fuzzy interval defined in this fashion has full membership in points b to c and
membership tends towards zero from b to a and from c to d.
There are several other membership functions such as Gaussian, Generalized Bell MF, Sigmoid MF, L-R MF etc. [Jang et al. (1997)]. However trapezoidal member-ship function is preferred because of its simplicity and computational efficiency. The quality of fuzzy partitioning depends on the choice of the values a, b, c, d which should be preferably provided by domain experts. When the attribute set is very large or in the absence of expertise, the parameter values can be obtained in an unsupervised manner making use of clustering techniques, where cluster centroids define the structure of data.

Interestingness Measures for Fuzzy Association Rules

Several interestingness measures are used to filter out the right set of actionable association rules from the larger set of rules mined by Association rule miningalgorithms. Normally support and confidence are most basic and essentially used measures in literature and were proposed by [Agrawal et al. (1993)]. The quality measures for fuzzy association rules has been described by [Dubois et al. (2003)], [Dubois et al. (2006)], [Burda (2014)], [Burda (2015)] in detail. The support measure defines statistical significance or usefulness of rule. A rule with support greater than user defined threshold is considered useful as any action taken based on this rule can give significant result. An itemset with high support that is greater than minimum threshold value is called as frequent or large itemset. It possesses an important downward closure property or Apriori property which states that all subsets of a frequent itemset are also frequent.

Fuzzy Cardinality

For binary attributes, support is an important measure that is used in validating generated association rules as well as in defining other interestingness measures. In boolean transactions, item can be either present or absent, hence support count is defined in terms of frequency of occurrence of an item set as given in section 2.For comparing support across data sets, relative support is used which is obtained by dividing support count by cardinality of dataset. The cardinality of dataset is number of transactions in dataset D and that is also the maximum possible support. The definitions of fuzzy support and fuzzy cardinality need to be suitably extended. In transaction dataset with fuzzy attributes, maximum possible support need not to be equal to number of transactions as maximum membership value for attribute term pair in a transaction may be less than 1. Hence for fuzzy data, cardinality needs to be replaced by maximum possible value of membership function. In case of quantified attributes, attributes gets partitioned into set of fuzzy linguistic terms and support count need to be defined for each such attribute term pair representing the corresponding linguistic term.

Definition : (Fuzzy Support Count (A)): For any attribute A partitioned into m fuzzy partitions defined as above, the support count of attribute A is defined as S(A) = $\max_{1\leq j\leq m} S(F^{j})$

Definition: (Fuzzy Cardinality of dataset D): For a dataset D with n transactions

and fuzzy attribute set X, each having its corresponding term sets, the fuzzy cardi-nality n_F is defined as $n_F = \max_{A\epsilon X}(S(A))$

To illustrate the above definitions, following dataset Employee is considered with

quantified attributes Age and Income, that are partitioned into fuzzy partitions. Attribute Age is partitioned into three linguistic terms {Young, Middle, Senior}, and the attribute Income is partitioned into {Low, Medium, High}. Membership values are computed and are listed in Table 1.

From the Table 1 support count for each attribute term pair can be computed

Table 1. Membership Values For Attributes Age And Income

	Age			Income		
Trans. No.	Young	Middle	Senior	Low	Medium	High
1	0.022175	0.977825	0	0	0	1
.	.	.	.	.	.	.
1195	0	0.013839	0.986161	0.197432	0.802568	0
1196	0.737482	0.262518	0	0.208998	0.791002	0
1197	0	0.976534	0.023466	0.23213	0.76787	0
1201	0	0	1	0.243697	0.756303	0

1202	0.451359	0.548641	0	0.255263	0.744737	0
1203	0.737482	0.262518	0	0.255263	0.744737	0
1204	0	1	0	0.266829	0.733171	0
1205	1	0	0	0.289961	0.710039	0
.	.	.	.	.	.	.
2920 Σ	1	0	0	1	0	0
	956.1073	1196.702	767.1907	**1681.304**	167.8439	1070.852

using the definitions provided above by adding membership values in a column. Support-Count of $\langle Age, Young\rangle =$ 956.1073
Support(Age) = Maximum {956.1073, 1196.702, 767.1907}= 1196.702
Support(Income) = 1681.304
and Cardinality of Employee Dataset is

= Maximum $\{Support(Age), Support(Income)\}$

= Maximum {1196.702, 1681.304} = 1681.304

Definition : (Fuzzy Support Count of Fuzzy Association Rule): For a dataset Dwith n transactions and any fuzzy linguistic attributes A and B, the support count

j

of fuzzy association rule (A, T^{i}) →(B, T) is defined as follows:

$F\_uzzy\ Support\ Count((A,T^{i})\rightarrow(B,T^{j}))$ $\underline{A \qquad B}$

n_F

In the above example,
Support-Count of rule $\langle Age, Middle\rangle \rightarrow \langle Income, High\rangle =$ 539.0905

Thus above definitions can be used to compute support count of fuzzy association rules.

Fuzzy Support Matrix

Many researchers have used 2 2 contingency table to define interestingness mea- sures ([Tan et al. (2004)], [Zembowicz and Ż ytkow (1996)], [Lenca et al. (2007)], [Yao and Zhong (1999)], [Xuan-hiep et al. (2006)]). The table actually stores fre- quency or support count of sets $A\overline{B}, \overline{A}B, \overline{A}\,\overline{B}$ and AB. These support counts can be easily computed for binary attributes as it is either present or absent. However for quantified or categorical attributes fuzzy approach can be used to handle this vagueness. A small or negligible value of quantified attribute may define absence and a high or substantial value may represent strong presence.

For quantified attributes, computing support count of A $\overline{B}$ is equivalent to comput-

ing the support count of rule $A \rightarrow \overline{B}$. Suppose, each attribute A is partitioned into fuzzy sets $F^1, F^2, .F^m$ where m ≥2, and is represented by m linguistic variables.

Since these fuzzy sets represent intervals, there is some order and $F^{\,1}$ is usually la- belled low indicating negligible presence or absence of A and F^m is usually labelled High to represent substantial presence of A. The support count of F^1 or F^{Low} is

A A

equivalent to support count of $\overline{A}$ and support of F^m or F^{High} is equivalent to sup-

A A

port count of A.

For attributes A and B the fuzzy association rule $A_{Low} \rightarrow B_{Low}$ indicates when ”A is in low quantity so is B” and is equivalent to negative implication of $\overline{A} \rightarrow \overline{B}$. The support count of $\overline{A} \rightarrow \overline{B}$ can be replaced by fuzzy support count of $A_{Low} \rightarrow B_{Low}$. The fuzzy association rule $A_{Low} \rightarrow B_{High}$ implies when A is in low quantity then B is in high quantity, thus approximates $\overline{A} \rightarrow B$ hence support count of $\overline{A} \rightarrow B$ can be replaced by fuzzy support count of $A_{Low} \rightarrow B_{High}$. Similarly the fuzzy supportof $A_{High} \rightarrow B_{Low}$ approximates $A \rightarrow \overline{B}$. The fuzzy association rule $A_{High} \rightarrow B_{High}$ implies A is high so is B indicating positive relationship $A \rightarrow B$.

The support matrix is defined for two attributes A and B. Each attribute is par- titioned into m fuzzy partitions $\{F^{\,1}, F^{\,2}, F^{\,3}, ...F^m\}$ m≥2 where $F^{\,1}$ indicates Low and F^m indicates high giving rise to fuzzy sets $\{F^1, F^2, F^3, ...F^m\}$, A A A A

which for attribute A are equivalent to $\{F^{Low}, ..., F^{High}\}$. The fuzzy sets for

A

attribute B will be $\{F^{Low}, ..._B\, T^{High}\}._B$ Let $S^{AB}\, A_{LH}$

denote fuzzy support count for

$A_{Low} \rightarrow B_{High}$, where L and H are abbreviations for Low and High respectively. The term S^{AB}_{L+} indicates support count when attribute A is Low while attributes B can take any possible value from low to high. The term S^{AB}_{+L} indicates support count when attribute A takes any possible values while attribute B is low. Similar
interpretation can be given for the terms S^{AB}_{H+} and S^{AB}_{+H}. The support matrix is

Table 2. Fuzzy Support Matrix

Attributes	B_{Low}		B_{High}	
A_{Low}	S^{AB}_{LL}	....	S^{AB}_{LL}	S^{AB}_{L+}
....	....	....	....	.....
A_{High}	S^{AB}_{HL}	....	S^{AB}_{HH}	S^{AB}_{H+}
	S^{AB}_{+L}	....	S^{AB}_{+H}	n_f

Table 3. Fuzzy Support Matrix For Three Linguistic Variables

Attributes	B_{Low}	B_{Medium}	B_{High}	
A_{Low}	S^{AB}_{LL}	S^{AB}_{LM}	S^{AB}_{LL}	S^{AB}_{L+}
A_{Medium}	S^{AB}_{ML}	S^{AB}_{MM}	S^{AB}_{MH}	S^{AB}_{M+}
A_{High}	S^{AB}_{HL}	S^{AB}_{HM}	S^{AB}_{HH}	S^{AB}_{H+}
	S^{AB}_{+L}	S^{AB}_{+M}	S^{AB}_{+H}	n_f

defined as in Table 2.

For simplicity and uniformity we can choose m = 3 with three fuzzy partitions labeled Low, Medium and High. The middle partition separates the two extreme partitions Low and High. In this case the support matrix will be in the form givenin Table 3. The above definitions are illustrated with some examples of computed support matrices for available datasets.

Conclusion

The fuzzy support matrix has been used for defining various interestingness mea- sures for quantified and ordered categorical attributes in a uniform and consistent manner. This extends the concept of Complimentary and Substitute attributes using the Odds Ratio as an interestingness measure. A new measure is defined to assess the relevance between two attributes. One important applications of substitute at- tributes is proposed to reduce dimensions by combining two substitute attributes and by removing irrelevant attributes. Moreover, there is a need to explore other possibilities of applying these concepts in various domains as well.

The extended definition of existing measures are applied to identify new attribute characteristics. Theoretical evaluation of various properties for interestingness mea- sures has been provided both in structural and behavioral aspects. A set of eight properties are considered for evaluating interestingness measures. Clustering based grouping is applied which helps in identifying a representative set of eight measures. Experimental study of these eight measures on different datasets is required to further strengthen their importance.

References

AGRAWAL, R., IMIELIŃSKI, T., AND SWAMI, A. 1993. Mining association rules between sets of items in large databases. In *Acm sigmod record*. Vol. 22. ACM, 207–216.

BAYARDO JR, R. J. AND AGRAWAL, R. 1999. Mining the most interesting rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 145–154.

BRIN, S., MOTWANI, R., AND SILVERSTEIN, C. 1997. Beyond market baskets: Generalizing association rules to correlations. In *Acm Sigmod Record*. Vol. 26. ACM, 265–276.

BURDA, M. 2014. Interest measures for fuzzy association rules based on expectations of independence. *Advances in Fuzzy Systems 2014*, 2.

BURDA, M. 2015. Lift measure for fuzzy association rules. In *Strengthening Links Between Data Analysis and Soft Computing*. Springer, 249–260.

CHAN, K. C. AND AU, W.-H. 1997. Mining fuzzy association rules. In *Proceedings ofthe sixth international conference on Information and knowledge management*. ACM, 209–215.

DIEU, P. D. Logic in knowledge systems. *Faculty of Technology, Hanoi National* .

DOUGHERTY, J., KOHAVI, R., SAHAMI, M., ET AL. 1995. Supervised and unsupervised discretization of continuous features. In *Machine learning: proceedings of the twelfth international conference*. Vol. 12. 194–202.

DUBOIS, D., HÜLLERMEIER, E., AND PRADE, H. 2003. A note on quality measures for fuzzy association rules. In *International Fuzzy Systems Association World Congress*. Springer, 346–353.

DUBOIS, D., HÜLLERMEIER, E., AND PRADE, H. 2006. A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery 13*, 2, 167–192.

GENG, L. AND HAMILTON, H. J. 2006. Interestingness measures for data mining: A survey.

ACM Computing Surveys (CSUR) 38, 3, 9.

HILDERMAN, R. J. AND HAMILTON, H. J. 1999. *Knowledge discovery and interestingness measures: A survey*. Citeseer.

HOLEŇA, M. 2009. Measures of ruleset quality for general rules extraction methods. *In- ternational Journal of Approximate Reasoning 50,* 6, 867–879.

JANG, J.-S. R., SUN, C.-T., AND MIZUTANI, E. 1997. Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence.

KAMBER, M. AND SHINGHAL, R. 1996. Evaluating the interestingness of characteristic rules. In *KDD*. 263–266.

KLEMENT, E. P., MESIAR, R., AND PAP, E. 2004. Triangular norms. position paper ii: general constructions and parameterized families. *Fuzzy Sets and Systems 145,* 3, 411–438.

LENCA, P., VAILLANT, B., MEYER, P., AND LALLICH, S. 2007. Association rule interest-ingness measures: Experimental and theoretical studies. In *Quality Measures in Data Mining*. Springer, 51–76.

LENT, B., SWAMI, A., AND WIDOM, J. 1997. Clustering association rules. In *Data Engi- neering, 1997. Proceedings. 13th International Conference on*. IEEE, 220–231.

MCCALL, S. 1967. *Polish Logic, 1920-1939*. OUP Oxford.

MILLER, R. J. AND YANG, Y. 1997. Association rules over interval data. *ACM SIGMOD Record 26,* 2, 452–461.

OLADIPUPO, O. O. 2012. a fuzzy association rule mining expert-driven approach to knowl- edge acquisition. Ph.D. thesis, Covenant University.

PIATETSKY-SHAPIRO, G. 1991. Discovery, analysis, and presentation of strong rules.

Knowledge discovery in databases, 229–238.

records. *Computer 43,* 10, 77–81.

RAMDASI, S. AND SHIRWAIKAR, S. 2016. Interpretability of fuzzy clusters by fuzzy associ-ation rules using cluster based fuzzy partitioning.

RUSNOK, P. AND BURDA, M. 2017. Global quality measures for fuzzy association rulebases. In *Advances in Fuzzy Logic and Technology 2017*. Springer, 268–276.

SAVASERE, A., OMIECINSKI, E. R., AND NAVATHE, S. B. 1995. An efficient algorithm for mining association rules in large databases. Tech. rep., Georgia Institute of Technology.

SCHWEIZER, B. 1991. Thirty years of copulas. In *Advances in probability distributions withgiven marginals*. Springer, 13–50.

SRIKANT, R. AND AGRAWAL, R. 1995. Mining generalized association rules.

Swati R. Ramdasi; Shailaja C. Shirwaikar; Vilas Kharat
International Journal of Fuzzy Computation and Modelling (IJFCM), Vol. 2, No. 4, 2019

TAN, P.-N., KUMAR, V., AND SRIVASTAVA, J. 2002. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD interna- tional conference on Knowledge discovery and data mining*. ACM, 32–41.

TAN, P.-N., KUMAR, V., AND SRIVASTAVA, J. 2004. Selecting the right objective measure for association analysis. *Information Systems 29,* 4, 293–313.

XUAN-HIEP, H., FABRICE, G., AND HENRI, B. 2006. Discovering the stable clusters between interestingness measures.

YANG, Y. AND SINGHAL, M. 1999. Fuzzy functional dependencies and fuzzy association rules. In *International Conference on Data Warehousing and Knowledge Discovery*. Springer, 229–240.

YAO, Y. AND ZHONG, N. 1999. An analysis of quantitative measures associated with rules. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 479– 488.

ZADEH, L. A. 1965. Fuzzy sets. *Information and control 8,* 3, 338–353.

ZEMBOWICZ, R. AND Ż YTKOW, J. M. 1996. From contingency tables to various forms of knowledge in databases. In *Advances in knowledge discovery and data mining*. Ameri-can Association for Artificial Intelligence, 328–349.