Overview of Fuzzy Associations Mining[1]

Guoqing Chen*, Qiang Wei*, Etienne Kerre**, Geert Wets***
*School of Economics and Management, Tsinghua University, Beijing 100084, China
**Dept. of Applied Mathematics and Computer Sciences, University of Gent, 9000 Gent, Belgium
***Faculty of Applied Economic Sciences, Limburg University, 3590 Diepenbeek, Belgium

Abstract –Associations, as specific forms of knowledge, reflect relationships among items in databases, and have been widely studied in the fields of knowledge discovery and data mining. Recent years have witnessed many efforts on discovering fuzzy associations, aimed at coping with fuzziness in knowledge representation and decision support processes. This paper focuses on associations of three kinds, namely, association rules, functional dependencies and pattern associations, and overviews major fuzzy logic extensions accordingly.

I. INTRODUCTION

Data mining, also called knowledge discovery in databases, is regarded as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable knowledge in large-scale data. Of particular interest in this paper is the discovery of associations that reflect relationships among items in databases. Generally speaking, associations may be categorized into several kinds, such as association rules (ARs), functional dependencies (FDs), and pattern associations (PAs).

Usually, associations of a typical kind are ARs, which have been extensively investigated in the field. In general, an AR $X \Rightarrow Y$ expresses the semantics that "occurrence of X is associated with occurrence of Y", where X and Y are collections of data items. Dsupport($X \Rightarrow Y$) and Dconfidence($X \Rightarrow Y$) could be used to evaluate the strength and certainty of $X \Rightarrow Y$, respectively. Since Agrawal et al. (1993) introduced the notion of (Boolean) ARs, mining ARs has attracted many research efforts along with a large number of AR applications.

In addition to association rules, FD is another kind of association of interest. FD is an important notion in relational database modeling (Codd, 1970; Chen, 1998). Generally speaking, a FD X→Y states that values of Y are uniquely determined by values of X, where X and Y are collections of data items. Classically, FDs could be constructed logically. However, in the context of data mining as a type of reverse engineering, the discovery of FDs has received considerable attention (Castellanos & Saltor, 1993; Bell & Brockhausen, 1995; Huhtala & Karkkainen, 1998a, 1998b; Liao & Wang et al., 1999; Savnik & Flach, 2000; Bosc & Pivert et al., 2001; Wei & Chen et al., 2002), because numerous database applications have generated a huge amount of data stored in distributed environments and with diversified structures, where many FDs might not originally be known or thought of being important, or have been hidden, but may be useful and interesting.

Finally, pattern associations are a third kind of associations. Discovering the relationships among time-series data (e.g., stocks, sales) is of particular interest since the time-series patterns reflect the evolution of changes in data values with sequential factors such as time. For instance, an example of such a case is "Firm *A*'s IT expenditure pattern is similar/opposite to Firm *B*'s IT expenditure pattern" in the context of IT organizational learning/diffusion. Apparently, it will be useful to discover such pattern associations.

However, it will be shown that in many situations discovering associations involves uncertainty and imprecision, particularly fuzziness. The necessity of applying fuzzy logic in data mining is twofold: one is

[1] Partly supported by China's National Natural Science Foundation (79925001/70231010), and the Bilateral Scientific & Technological Cooperation Programmes between China and Flandres/Czech. Cooresponding author: Guoqing Chen, chengq@em.tsinghua.edu.cn

that fuzziness is inherent in many problems of knowledge representation, and the other is that high-level managers or complex decision processes often deal with generalized concepts and linguistic expressions, which are generally fuzzy in nature. Typically, “sharp boundaries” and “partial belongings” are two main problems encountered in association mining. Moreover, fuzziness may prevail in many other association cases in which imprecision, matching, similarity, implication, partial truth or the like is present. As indicated already, existing efforts on fuzzy extensions can be distinguished into three main streams, namely, fuzzy AR (FAR), fuzzy/partial satisfied FDs (FFD/FD_d), and fuzzy logic in PA (FPA).

II. FUZZY LOGIC IN QUANTITATIVE AR

Though Boolean ARs are meaningful, there are many other situations where data items concerned are usually categorical or quantitative. Srikant et al. (1996) presented an approach to discover quantitative ARs by transferring quantitative items into binary items by partitioning continuous domains. For example, if item Age in database D takes values from (0, 100], then it could be partitioned into three new items such as Age(0,30], Age(30,60], and Age(60,100], respectively. Then a new database D’ is constructed. Differently from Boolean ARs, quantitative ARs represents “Quantity of X is associated with Quantity of Y”.

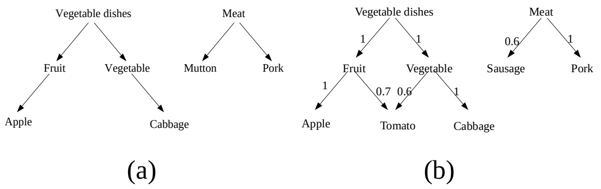
Apparently, whatever partitioning methods are applied (Srikant & Agrawal, 1996; Mazlack, 2000), “sharp boundaries” remains a problem, which may under-estimate or over-emphasize the elements near the boundaries, and may therefore lead to an inaccurate representation of semantics.

In dealing with “sharp boundaries” problem, fuzzy sets and fuzzy items, usually in forms of labels or linguistic terms, are used and are defined onto the domains (Fu et al., 1998; Mazlack, 2000; Chien & Lin et al., 2001; Gyenesei, 2001). For example, for Age, some fuzzy items along with corresponding fuzzy sets may be defined on its domain U_{Age} such as Young, Middle and Old, which will be used to constitute a new database *D”* with partial belongings of original item values (in *D*) to each of the new items (in *D”*). Several attempts have been made in defining fuzzy sets onto continuous domains (Fu et al., 1998; Gyenesei, 2000a; Shu et al., 2000; Chien & Lin et al., 2001).

With the above extended database *D”*, conventional notions of Dsupport/Dconfidence could be extended as well. Though a few measures have been proposed, they are in a similar spirit that Σcount operator is used for fuzzy cardinality. Subsequently, with these extended measures incorporated, several mining algorithms have been proposed (Lee & Hyung; 1997; Kuok et al., 1998; Hong & Kuo, 1999a, 1999b; Gyenesei, 2000a, 2001; Shu & Tsang et al., 2001; Chan & Au, 2001).

III. FUZZY AR WITH FUZZY TAXONOMIES

Srikant & Agrawal (1995) presented a method to discover the so-called generalized AR based on concept taxonomies as shown in Figure 1 (a). So, “Fruit⇒Meat”, rather than “Pork⇒Apple”, is more general and has more potential to be discovered.



(a) (b)

Figure 1. Exact Taxonomies and Fuzzy Taxonomies.

In 1999, Wei & Chen extended generalized AR with fuzzy taxonomies, by which partial belongings could be incorporated. For example, given fuzzy taxonomies in Figure 1(b), Tomato belongs to Fruit and Vegetable with different degrees respectively, which may be semantically meaningful. More concretely, some corresponding fuzzy extensions on measures and mining methods are proposed to fit the fuzzy context (Chen & Wei, 1999, 2002; Wei & Chen, 1999).

Furthermore, a recent effort has been made as described in Chen & Wei et al. (1999, 2002), which presents an approach to incorporate linguistic hedges on existing fuzzy taxonomies. Then after applying all the proper hedges in a given linguistic pool *H* onto the items, new fuzzy taxonomies with all modified items could be derived, as shown in Figure 2. In so doing, the problem of mining linguistic association rules with hedges pool *H* on fuzzy taxonomies could be transferred to the problem of mining fuzzy association rules on the new taxonomies.

Moreover, some new optimizations have been incorporated into mining process (Wei & Chen et al., 2000), which could avoid item exploration while doing linguistic modification.

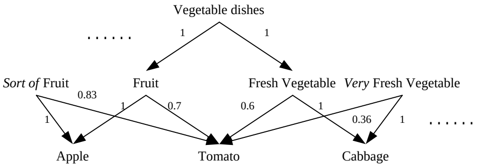


Figure 2. Linguistically Modified Fuzzy Taxonomies

IV. OTHER FUZZY EXTENSIONS

In addition to the above two major directions, several fuzzy extensions have been made on interestingness measures. Hullermeier (2001b) proposed a measure called interestingness degree, which could be seen as the increase in probability of Y caused by the occurrence of X. A few attempts were to introduce thresholds for filtering databases in dealing with very low membership degrees (Lee & Hyung, 1997; Kuok & Fu et al., 1998; Hullermeier, 2001b). Additionally, some other efforts focused on constructing similar measures to conventional Dsupport/Dconfidence (Shragai & Scgreider, 2001; Gyenesei & Teuhola, 2001). Another attempt to mention is the work by Au & Chan (1997, 1998), who proposed a certainty measure, called adjusted difference, based on statistical test.

Sometimes, one may think of users paying more attention to certain attributes than to others. Similarly, in fuzzy AR mining, weights could also be applied to distinguish the importance of different items. Some approaches by Cai & Fu et al. (1998), Gyenesei (2000b), Shu & Tsang et al. (2000) etc. have already been proposed, which are basically similar.

Notably, the number of filtering thresholds and weights as well as their determination may be an issue of concern.

V. FUZZY IMPLICATION BASED ARs

As indicated previously, traditionally a rule of X⇒Y is referred to as a relationship between X and Y and modeled by conditional probability (e.g., Dconfidence) for X-to-Y. In further investigating X-to-Y relationships, a more logic-oriented view may be taken so as to reflect, to certain extent, implication from X to Y. Still in terms of ARs and in fuzzy contexts, a few efforts have been made to consider partial degrees that X implies Y. For instance, in (Chen & Wei et al., 1999; Dubois & Hullermeier et al., 2001; Hullermeier, 2001a), fuzzy implication is introduced to represent the degree that a tuple supports X-to-Y. Fuzzy Implication Operators (FIOs) are used to express the logic inference semantics. Since FIO is generally not symmetric, X⇒Y and Y⇒X could be distinguished. More concretely, Hullermeier (2001a) discussed several different type of fuzzy implication based ARs theoretically.

Recently, Chen & Yan et al. (2002) introduced a notion called degree of implication (denoted as Dimp) to evaluate the strength of ARs from a more logic-oriented viewpoint. In so doing, a proper selection of FIO and t-norm combinations could help avoid database scanning, and therefore improve the efficiency of ARs generation.

VI. MINING FDs WITH UNCERTAINTIES

This section focuses on discovering fuzzy FDs (FFD), and discovering FDs with partial degrees (FD_d), respectively.

First, fuzzy functional dependencies (FFD) are extensions of classical FD, aimed at dealing with fuzziness in databases and reflecting the semantics that *close* values of a collection of items are dependent on *close* values of a collection of different items. Generally, FFDs have different forms, depending on the different aspects of integrating fuzzy logic in classical FDs.

Somewhat differently from the ways that are of a typical data mining nature, Cubero et al. (1995, 1999) presented a method of data summarization through FFDs in both crisp and fuzzy databases, in which projection operations are applied to reduce the amount of data in databases without loss of information. Recently, Wang et al. (2002) presented a method to discover FFDs in similarity-based databases with an incremental strategy.

Generally speaking, the discovered FFDs expressed the semantics that "similar Xs infer to similar Ys" to some extent. Moreover, Yang & Singhal (2001) attempted to present a framework of linking FFDs and FARs in a closer manner.

In massive databases where noisy or incomplete/imprecise information exists, classical FDs may be too restrictive to hold, since the correspondence of equal X-Y values must be 100% satisfied, by definition. However, it may be meaningful to take into account partial satisfaction of FD, being capable of tolerating the noisy or incomplete/imprecise information at certain degrees.

Huhtala et al. (1998a, 1998b) have explored a notion called approximate FD so as to represent FD that "almost holds". Recently, Wei & Chen et al. (2002) presented the notion of FD with degree of satisfaction (FD_d), which is another measure for degree that a FD holds in D. Further, Wei et al. (2003) constructed Armstrong-Analogous Axioms, based on which a minimal set of qualified $(FDs)_d$ could be derived efficiently.

VII. FUZZY LOGIC IN PATTERN ASSOCIATIONS

Discovering relationships among time series is of particular interest since time series patterns reflect the evolution of changes in item values with sequential factors, e.g., time. The value evolution of each time series item is viewed as a pattern over time, and the similarity between any two patterns is measured by pattern matching.

Concretely, two major issues are involved in dealing with similar time series patterns. One is the measurement for pair-wise similarities. The problems related to this issue center around how to define the difference between any two patterns, say, in terms of "distance" and how to match the series in points of time. The other issue is the grouping of the similar patterns, in which fuzzy relations and clustering may play an important role. Usually, static similarities relationship are studied, which could be obtained by computing the "distance" pair-wisely in a fixed matching fashion as shown in Figure 3. In this case, the matching scheme for curves *a* and *b* cannot be applied to the matching between curves *b* and *c*; and vice versa. Thus, any pair of curves *a*, *b* and *c* reflects a certain matching scenario, which is static schematically.

Furthermore, the way to discover the similarities among the curves could be improved by matching the patterns dynamically. This can be done by using the *Dynamic Time Warping* (*DTW*) method, a method used in speech recognition (Berndt & Clifford, 1996). Chen & Wei et al. (2001) presented a method based on *DTW* to discover pattern associations.

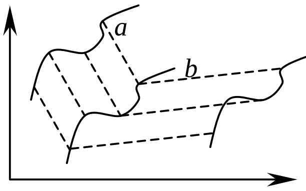


Figure 3. Static Matching Schemes.

Finally, it is worthwhile to indicate that, though at the inception stage, discovering pattern associations is deemed a promising area of theoretical and practical explorations and many attempts are expected to emerge, in that fuzzy logic will play an important role.

VIII. CONCLUDING REMARKS

This paper has aimed at providing readers with a brief overview on discovering fuzzy associations. Discussions have centered around fuzzy ARs in dealing with partitioning quantitative data domains; crisp taxonomic belongings and linguistically modified rules; various fuzzy mining measures from different perspectives such as interestingness and logic implication; fuzzy/partially satisfied FDs for handling data closeness and noise tolerance; and time-series data patterns that are similar with partial degrees. A more complete overview with details is to appear in a separate paper.

REFERENCES

[1] Agrawal, R.; Imielinski, T.; Swarmi, A., 1993. Mining Association Rules between Sets of Items in Large Databases, In Proceedings of the ACM-SIGMOD 1993 International Conference on Management of Data, Washington D. C., U.S.A., pp. 207-216.

[2] Chen G. Q.; Wei, Q.; Kerre, E. E., 1999. *Fuzzy Data Mining: Discovery of Fuzzy Generalized Association Rules*, in Recent Research Issues on Management of Fuzziness in Databases, in the

Physica-Verlag series "Studies in Fuzziness and Soft Computing", Springer-Verlag, NY, U.S.A.

[3] Chen, G. Q.; Yan, P.; Kerre, E. E., 2002. Mining Fuzzy Implication-Based Association Rules in Quantitative Databases, Proceedings of FLINS2002, Belgium.

[4] Chen, G. Q., 1998. Fuzzy Logic in Data Modeling: semantics, constraints and database design, Kluwer Academic Publishers, Boston, MA, U.S.A.

[5] Chen, G. Q.; Wei, Q.; Liu, D.; Wets, G.., 2002. Simple Association Rules (SAR) and the SAR-Based Rule Discovery, Journal of Computer & Industrial Engineering 43 (2002), 721-733.

[6] Chen, G. Q.; Wei, Q., 2002. Fuzzy Association Rules and the Extended Mining Algorithms, Information Sciences, 147, pp. 201-228.

[7] Chen, G. Q.; Wei, Q.; Zhang, H., 2001. Discovering Similar Time-Series Patterns with Fuzzy Clustering and DTW Methods, IFSA/NAFIPS2001, Vancouver, BA, Canada.

[8] Chien, B. C.; Lin, Z. L.; Hong, T. P., 2001. An Efficient Clustering Algorithm for Mining Fuzzy Quantitative Association Rules, in Proceedings of the 9th International Fuzzy Systems Association World Congress, July 25-28, Vancouver, Canada, pp. 1306-1311.

[9] Codd EF, 1970. A Relational Model for Large Shared Data Banks. Communications of the ACM, 13(6): 377-387.

[10] Cubero, J. C. et al, 1999. Data Summarization in Relational Databases through Fuzzy Dependencies, Information Sciences, Vol. 121 (3-4), pp.233-270.

[11] Cubero, J.C.; Medina, J. M.; Pons, O.; Vila, M.A., 1995. *Rules discovery in fuzzy relational databases*. In Conference of the North American Fuzzy Information Processing Society, NAFIPS'95. Maryland (USA). IEEE Computer Society Press, pp. 414-419.

[12] Dubois, D.; Hullermeier, E.; Prade, H., 2001. Toward the Representation of Implication-Based Fuzzy Rules in Terms of Crisp Rules, in Proceedings of IFSA/NAFIPS2001, Vancouver, BA, Canada.

[13] Fu, A. et al., 1998. Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes, in Proceedings of 1st Intl. Symposium on Intelligent Data Engineering and Learning (IDEAL'98), pages 263--268.

[14] Gyenesei, A., 2000a. *A fuzzy approach for mining quantitative association rules*, TUCS technical reports 336, University of Turku, Department of Computer Science, Lemminkisenkatu14, Finland

[15] Gyenesei, A., 2000b. *Mining Weighted Association Rules for Fuzzy Quantitative Items.* In Proceedings of PKDD Conference, September 13-16, 2000, Lyon, France. pp. 416-423.

[16] Gyenesei, A., 2001. Fuzzy Partitioning of Quantitative Attribute Domains by a Cluster Goodness Index, http://citeseer.nj.nec.com/440030.html.

[17] Gyenesei, A.; Teuhola, J., 2001. Interestingness Measures for Fuzzy Association Rules, PKDD 2001: Freiburg, Germany, pp. 152-164.

[18] Hong, T. P.; Kuo, C. S.; Chi, S. C., 1999a. A fuzzy data mining algorithm for quantitative values, The Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, pp. 480-483.

[19] Hong, T. P.; Kuo, C. S.; Chi, S. C., 1999b. Mining association rules from quantitative data, Intelligent Data Analysis, Vol. 3, No. 5, pp. 363-376.

[20] Huhtala, Y.; Karkkainen, J.; Paokka, P.; Toivonen, H., 1998a. TANE: An Efficient Algorithm for Discovering Functional and Approximate Dependencies, http://citeseer.nj.nec.com/huhtala99tane.html.

[21] Huhtala, Y.; Karkkainen, J.; Porkka, P.; & Toivonen, H., 1998b. *Efficient Discovery of Functional and Approximate Dependencies Using Partitions.* Proc. 14th Int. Conf. on Data Engineering, IEEE Computer Society Press.

[22] Hullermeier, E., 2001a. Implication-Based Fuzzy Association Rules, ECML/PKDD 2001, Freiburg, Germany.

[23] Hullermeier, E., 2001b. Fuzzy Association Rules: Semantics Issues and Quality Measures, http://citeseer.nj.nec.com/.

[24] Kuok, C. M.; Fu, A.; Wong, M. H., 1998. *Mining Fuzzy Association Rules in Databases*, SIGMOD Record, pp. 41-46, Vol. 27, No. 1.

[25] Lee, J. H.; Hyung, L. K., 1997. An Extension of Association Rules using Fuzzy Sets, Seventh IFSA World Congress, Prague, pp. 399-402.

[26] Liao, S. Y.; Wang, H. Q.; Liu, W. Y., 1999. Functional Dependencies with Null Values, Fuzzy Values, and Crisp Values, IEEE Transactions on Fuzzy Systems, Vol. 7, No. 1, pp. 97-103.

[27] Mazlack, L. J., 2000. Approximate Clustering in Association Rules, *19th International Conference of the North American Fuzzy Information Processing Society - NAFIPS 2000,* Atlanta, pp. 256-260.

[28] Savnik, I.; Flach, P. A., 2000. Discovery of Multi-valued Dependencies from Relations, report00135. http://citeseer.nj.nec.com/ savnik00discovery.html.

[29] Shragai, A.; Scgreider, M., 2001. Discovering Quantitative Association Rules in Database, http://citeseer.nj.nec.com/.

[30] Shu, J. Y.; Tsang, E. C. C.; Daniel; Yeung, S., 2001. Query Fuzzy Association Rules in Relational Database, Proceedings of IFSA/NAFIPS 2001, Vancouver, BA, Canada.

[31] Shu, J.; Tsang, E.; Yeung, D. S.; Shi, D., 2000. Mining fuzzy association rules with weighted items, In: Proc. IEEE Int'l Conf. on System, Man and Cybernetics (SMC2000), Nashville, Tennessee.

[32] Srikant, R.; Agrawal, R., 1995. *Mining Generalized Association Rules*, in Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland.

[33] Srikant, R.; Agrawal, R., 1996. *Mining Quantitative Association Rules in Large Relational Tables*, SIGMOD'96 6/96 Montreal, Canada.

[34] Wang, S. L.; Shen, J. W.; Hong, T. P., 2002. Incremental discovery of functional dependencies based on partitions, Intelligent Data Analysis (in revision).

[35] Wei, Q.; Chen, G. Q., 1999. *Mining Generalized Association Rules with Fuzzy Taxonomic Structures*, in 18th Int'l Conf. of NAFIPS, New York, NY, USA, 477-481.

[36] Wei, Q.; Chen, G. Q., 2000. Association Rules with Opposite Items in Large Categorical Database, FQAS2000. Warsaw, Poland.

[37] Wei, Q.; Chen, G.Q., 2003. Mining a Minimal Set of Functional Dependencies with Degrees of Satisfaction, in Proc. of Intl. Conf. on Fuzzy Information Processing (FIP2003), Beijing China

[38] Wei, Q.; Chen, G.. Q.; Kerre, E. E., 2002. Mining Functional Dependencies with Degrees of Satisfaction in Databases, in Proceedings of Joint Conference on Information Sciences, Durham, NC, USA.

[39] Yang Y. P.; Singhal, M., 2001. Fuzzy Functional Dependencies and Fuzzy Association Rules, http://citeseer.nj.nec.com/.