Mining Gradual Dependencies based on Fuzzy Rank Correlation

Hyung-Won Koh and Eyke Hüllermeier[1]

Abstract We propose a novel framework and an algorithm for mining gradual dependencies between attributes in a data set. Our approach is based on the use of fuzzy rank correlation for measuring the strength of a dependency. It can be seen as a unification of previous approaches to evaluating gradual dependencies and captures both, qualitative and quantitative measures of association as special cases.

1 Introduction

In association analysis, a widely applied data mining technique, the goal is to find "interesting" associations in a data set, that is, dependencies between so-called itemsets (binary attributes) $\mathcal{A}$ and $\mathcal{B}$ expressed in terms of rules of the form "IF $\mathcal{A}$ THEN $\mathcal{B}$". The intended meaning of a rule of that kind is that, if $\mathcal{A}$ is present in a transaction, then $\mathcal{B}$ is likely to be present, too. Association rule mining has also been extended to the fuzzy case, in which the presence of an item in a transaction is a matter of degree [7].

Another type of association rule, called *gradual dependency*, has been introduced in [10] and was further studied in [2, 11]. As explained in Section 2, the idea is to express dependencies, not between the presence or absence of attributes, but between the *change* of the presence of fuzzy items in a transaction. The contribution of this paper is a novel framework for mining gradual dependencies that is based on the use of fuzzy rank correlation as a measure of confidence (Section 3). This framework can be seen as a unification of previous approaches and captures both, qualitative and quantitative measures of association (Section 4). We also propose an algorithm for mining gradual dependencies and illustrate the method on a wine quality data set.

Department of Mathematics and Computer Science, University of Marburg, Germany
{koh, eyke}@mathematik.uni-marburg.de

2 Gradual Dependencies

We adopt a feature-based representation of transactions (data records) and denote by $\mathbb{A}$ the (finite) set of underlying fuzzy attributes. Thus, each transaction is represented in terms of a feature vector $\boldsymbol{u}$, and for each $A \in \mathbb{A}$, $A(\boldsymbol{u}) \in [0,1]$ indicates the degree to which $\boldsymbol{u}$ has feature A or, say, to which A is present in $\boldsymbol{u}$. Correspondingly, the degree of presence of a feature subset $\mathcal{A} = \{A_1, \ldots, A_m\} \subset \mathbb{A}$, considered as a conjunction of primitive features $A_1, \ldots, A_m$, is given by $\mathcal{A}(\boldsymbol{u}) = \top(A_1(\boldsymbol{u}), A_2(\boldsymbol{u}), \ldots, A_m(\boldsymbol{u}))$, where $\top$ is a triangular norm (t-norm) serving as a generalized conjunction.

Given a data set consisting of N transactions $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$, a standard problem in (fuzzy) association analysis is to find all rules $\mathcal{A} \rightharpoonup \mathcal{B}$ whose support and confidence, defined as

$$\text{supp} = \sum_{i=1}^{N} \top(\mathcal{A}(\boldsymbol{u}_i), \mathcal{B}(\boldsymbol{u}_i)), \quad \text{conf} = \frac{\sum_{i=1}^{N} \top(\mathcal{A}(\boldsymbol{u}_i), \mathcal{B}(\boldsymbol{u}_i))}{\sum_{i=1}^{N} \mathcal{A}(\boldsymbol{u}_i)}, \tag{1}$$

exceed user-defined thresholds. A rule of such kind indicates the frequent occurrence of $\mathcal{B}$ given $\mathcal{A}$ (confidence), confirmed by sufficiently many examples (support). On a logical level, the meaning of a standard association rule $\mathcal{A} \rightharpoonup \mathcal{B}$ is captured by the material conditional. On a natural language level, such a rule is understood as an IF–THEN construct: If the antecedent $\mathcal{A}$ holds true, so does the consequent $\mathcal{B}$.

As mentioned above, another type of pattern, called *gradual dependency*, was introduced in [10]. Here, the idea is to express dependencies between the *direction of change* of attribute values. This idea is closely connected to so-called *gradual rules* in fuzzy logic. On a logical level, such rules are modeled in terms of residuated implication operators. Semantically, a rule $\mathcal{A} \rightharpoonup \mathcal{B}$ is often understood as "THE MORE the antecedent $\mathcal{A}$ is true, THE MORE the consequent $\mathcal{B}$ is true", for example "The larger an object, the heavier it is" [8]. This interpretation is arguable, however. In fact, to satisfy a gradual fuzzy rule in a logical sense, it is enough that $\mathcal{A}(\boldsymbol{u}) \leq \mathcal{B}(\boldsymbol{u})$; thus, there is actually no consideration of the change of an attribute value and, therefore, no examination of a tendency.

2.1 Evaluating Gradual Dependencies

Instead of pursing a logical approach using implication operators to evaluate a rule $\mathcal{A} \rightharpoonup \mathcal{B}$, it was proposed in [10] to take the so-called *contingency diagram* as a point of departure. A contingency diagram is a two-dimensional diagram in which every transaction $\boldsymbol{u}$ defines a point $(x, y) = (\mathcal{A}(\boldsymbol{u}), \mathcal{B}(\boldsymbol{u})) \in [0,1]^2$. Thus, for every transaction $\boldsymbol{u}$, the values on the abscissa and ordinate are

given, respectively, by the degrees $x = \mathcal{A}(\boldsymbol{u})$ and $y = \mathcal{B}(\boldsymbol{u})$ to which it satisfies the antecedent and the consequent part of a candidate rule.

Informally speaking, a gradual dependency is then reflected by the relationship between the points in the contingency diagram. In particular, a "THE MORE ... THE MORE" relationship manifests itself in an increasing trend, i.e., an approximate functional dependency between the x- and y-values: the higher x, the higher y tends to be. In [10], it was therefore suggested to analyze contingency diagrams by means of techniques from statistical regression analysis. For example, if a linear regression line with a significantly positive slope can be fit to the data, this suggests that indeed a higher $x = \mathcal{A}(\boldsymbol{u})$ tends to come along with a higher $y = \mathcal{B}(\boldsymbol{u})$.

A qualitative, non-parametric alternative to this numerical approach was proposed in [2]. Roughly speaking, to evaluate a candidate rule $\mathcal{A} \rightharpoonup \mathcal{B}$, the authors count the number of pairs of points (x, y) and (x', y') in the contingency diagram for which $x < x'$ and $y < y'$. As an advantage of this approach, note that it is more flexible in the sense of not making any assumption about the type of functional dependency; as opposed to this, the regression approach implicitly assumes a linear dependency. On the other hand, since the actual distances between the points are ignored, there is also a disadvantage, namely a loss of information about the strength of a relationship.

The two above approaches, the numerical and the qualitative one, essentially come down to looking for two types of correlation between the x- and y-values, namely the standard Pearson correlation and the rank correlation. The goal of this paper is to combine the advantages of both approaches. To this end, we propose to measure the strength of a dependency in terms of a *fuzzy rank correlation* measure that combines properties of both types of correlation. As will be seen, this measure is able to capture the strength of a tendency while remaining flexible and free of specific model assumptions. Our proposal is related to the approach presented in [12] but additionally offers a sound theoretical justification.

3 Fuzzy Rank Correlation

Consider $n \geq 2$ paired observations $\{(x_i, y_i)\}_{i=1}^n \subset (\mathbb{X} \times \mathbb{Y})^n$ of two variables X and Y, where $\mathbb{X}$ and $\mathbb{Y}$ are two linearly ordered domains; we denote $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. The goal of a rank correlation measure is to measure the dependence between the two variables in terms of their tendency to increase and decrease in the same or the opposite direction. If an increase in X tends to come along with an increase in Y, then the (rank) correlation is positive. The other way around, the correlation is negative if an increase in X tends to come along with a decrease in Y. If there is no dependency of either kind, the correlation is (close to) 0.

Several rank correlation measures are defined in terms of the number C of *concordant*, the number D of *discordant*, and the number N of *tied* data points. For a given index pair $(i, j) \in \{1, \ldots, n\}^2$, we say that (i, j) is concordant, discordant or tied depending on whether $(x_i - x_j)(y_i - y_j)$ is positive, negative or 0, respectively. A well-known example is Goodman and Kruskal's *gamma rank correlation* [9], which is defined as $\gamma = (C - D)/(C + D)$.

3.1 Fuzzy Equivalence and Order Relations

Bodenhofer and Klawonn [5] propose a fuzzy extension of the gamma coefficient based on concepts of fuzzy orderings and $\top$-equivalence relations, where $\top$ denotes a t-norm [3].

A fuzzy relation $E : \mathbb{X} \times \mathbb{X} \to [0, 1]$ is called *fuzzy equivalence* with respect to a t-norm $\top$, for brevity $\top$-equivalence, if it is reflexive ($E(x, x) = 1$), symmetric ($E(x, y) = E(y, x)$), and $\top$-transitive ($\top(E(x, y), E(y, z)) \leq E(x, z)$). Moreover, a fuzzy relation $L : \mathbb{X} \times \mathbb{X} \to [0, 1]$ is called *fuzzy ordering* with respect to a t-norm $\top$ and a $\top$-equivalence E, for brevity $\top$-*E-ordering*, if it is E-reflexive ($E(x, y) \leq L(x, y)$), $\top$-E-antisymmetric ($\top(L(x, y), L(y, x)) \leq E(x, y)$), and $\top$-transitive($\top(L(x, y), L(y, z)) \leq L(x, z)$). We call a $\top$-E-ordering L *strongly complete* if $\max(L(x, y), L(y, x)) = 1$ for all $x, y \in \mathbb{X}$. Finally, let R denote a strict fuzzy ordering associated with a strongly complete $\top$-E-ordering L; in the case of the well-known Łukasiewicz t-norm, defined by $\top(x, y) = \max(0, x + y - 1)$, this relation can simply be taken as $R(x, y) = 1 - L(x, y)$ [4].

3.2 The Fuzzy Gamma Rank Correlation

Consider a set of paired data points $\{(x_i, y_i)\}_{i=1}^n \subset (\mathbb{X} \times \mathbb{Y})^n$ and assume to be given two $\top$-equivalences $E_{\mathbb{X}}$ and $E_{\mathbb{Y}}$ and two strict fuzzy order relations $R_{\mathbb{X}}$ and $R_{\mathbb{Y}}$. Using these relations, the concepts of concordance and discordance of data points can be generalized as follows: Given an index pair (i, j), the degree to which this pair is concordant, discordant, and tied is defined, respectively, as

$$\tilde{C}(i, j) = \top(R_{\mathbb{X}}(x_i, x_j), R_{\mathbb{Y}}(y_i, y_j)), \tag{2}$$

$$\tilde{D}(i, j) = \top(R_{\mathbb{X}}(x_i, x_j), R_{\mathbb{Y}}(y_j, y_i)), \tag{3}$$

$$\tilde{T}(i, j) = \bot(E_{\mathbb{X}}(x_i, x_j), E_{\mathbb{Y}}(y_i, y_j)), \tag{4}$$

where $\top$ is a t-norm and $\bot$ is the dual t-conorm of $\top$ (i.e. $\bot(x, y) = 1 - \top(1 - x, 1 - y)$). The following equality holds for all index pairs (i, j):

$$\tilde{C}(i, j) + \tilde{C}(j, i) + \tilde{D}(i, j) + \tilde{D}(j, i) + \tilde{T}(i, j) = 1.$$

Adopting the simple sigma-count principle to measure the cardinality of a fuzzy set, the number of concordant and discordant pairs can be computed, respectively, as

$$\tilde{C} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{C}(i,j), \qquad \tilde{D} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{D}(i,j).$$

The *fuzzy ordering-based* gamma rank correlation measure $\tilde{\gamma}$, or simply "fuzzy gamma", is then defined as

$$\tilde{\gamma} = \frac{\tilde{C} - \tilde{D}}{\tilde{C} + \tilde{D}} \ . \tag{5}$$

From the definition of $\tilde{\gamma}$, it is clear that the basic idea is to decrease the influence of "close-to-tie" pairs (x_i, y_i) and (x_j, y_j). Such pairs, whether concordant or discordant, are turned into a partial tie, and hence are ignored to some extent. Or, stated differently, there is a smooth transition between being concordant (discordant) and being tied; see Fig. 1.

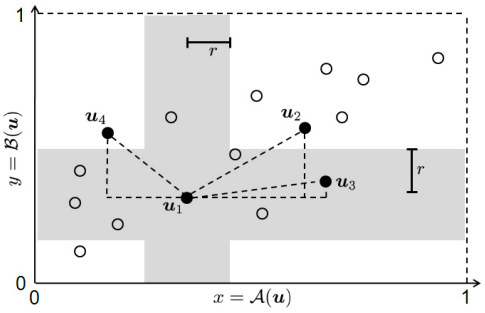

Fig. 1 Example of a contingency diagram. The pair $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ is concordant, while $(\boldsymbol{u}_1, \boldsymbol{u}_4)$ is discordant. Points with a distance $< r$ from $\boldsymbol{u}_1$ in one of the dimensions (gray region) are considered as partially tied with $\boldsymbol{u}_1$. For example, the pair $(\boldsymbol{u}_1, \boldsymbol{u}_3)$ is concordant to a degree < 1.

4 Mining Gradual Dependencies

Our idea is to evaluate a gradual dependency $\mathcal{A} \rightharpoonup \mathcal{B}$ in terms of two measures, namely the number of concordant pairs, $\tilde{C}$, and the rank correlation $\tilde{\gamma}$ as defined in (5). Comparing this approach with the classical setting of association analysis, $\tilde{C}$ plays the role of the support of a rule, while $\tilde{\gamma}$ corresponds to the confidence. These measures can also be nicely interpreted within the formal framework proposed in [7], in which every observation (in our case a pair of points $(\mathcal{A}(\boldsymbol{u}), \mathcal{B}(\boldsymbol{u}))$ and $(\mathcal{A}(\boldsymbol{v}), \mathcal{B}(\boldsymbol{v}))$) is considered, to a

certain degree, as an *example* of a pattern, as a *counterexample*, or as being *irrelevant* for the evaluation of the pattern. In our case, these degrees are given, respectively, by the degree of concordance, the degree of discordance, and the degree to which the pair is a tie.

4.1 Evaluation of Candidate Rules

More formally, we define the support and confidence of a gradual dependency $\mathcal{A} \rightharpoonup \mathcal{B}$ as follows:

$$\text{supp}(\mathcal{A} \rightharpoonup \mathcal{B}) = \tilde{C}, \quad \text{conf}(\mathcal{A} \rightharpoonup \mathcal{B}) = \frac{\tilde{C} - \tilde{D}}{\tilde{C} + \tilde{D}},$$

where

$$\tilde{C} = \sum_{\boldsymbol{u}_i} \sum_{\boldsymbol{u}_j} \tilde{C}(\boldsymbol{u}_i, \boldsymbol{u}_j) = \sum_{\boldsymbol{u}_i} \sum_{\boldsymbol{u}_j} \top \left(L\left(\mathcal{A}(\boldsymbol{u}_i), \mathcal{A}(\boldsymbol{u}_j)\right), L\left(\mathcal{B}(\boldsymbol{u}_i), \mathcal{B}(\boldsymbol{u}_j)\right)\right),$$
$$\tilde{D} = \sum_{\boldsymbol{u}_i} \sum_{\boldsymbol{u}_j} \tilde{D}(\boldsymbol{u}_i, \boldsymbol{u}_j) = \sum_{\boldsymbol{u}_i} \sum_{\boldsymbol{u}_j} \top \left(L\left(\mathcal{A}(\boldsymbol{u}_i), \mathcal{A}(\boldsymbol{u}_j)\right), L\left(\mathcal{B}(\boldsymbol{u}_j), \mathcal{B}(\boldsymbol{u}_i)\right)\right).$$

Considering the special case of the Łukasiewicz t-norm, it can be verified that $E(x, y) = [1 - |x - y|/r]_0^1$ is a $\top$-equivalence on $\mathbb{R}$ and $R(x, y) = [(x - y)/r]_0^1$ is a strict fuzzy ordering, where $[\cdot]_0^1$ denotes the mapping $a \mapsto \min(1, \max(0, a))$. Note that these relations are parameterized by the value $r \in (0, 1]$. For $r \to 0$, the confidence measure converges toward the classical (non-fuzzy) rank correlation, whereas for $r = 1$, we obtain $R(x, y) = x - y$ if $x \geq y$ and $= 0$ otherwise. The degree of concordance (discordance) is then proportional to the Euclidean distances, which means that this case is very close to the numerical evaluation in terms of Pearson correlation.

4.2 Rule Mining and Algorithmic Issues

Due to the associativity of a t-norm, the support of a rule $\mathcal{A} \rightharpoonup \mathcal{B}$ just corresponds to the support of the itemset $\mathcal{I} = \mathcal{A} \cup \mathcal{B}$. In other words, to compute a degree of concordance, there is no need to separate an itemset into an antecedent and a consequent part of a rule. Moreover, it is easy to see that the support measure is anti-monotone, i.e., $\text{supp}(\mathcal{I}) \leq \text{supp}(\mathcal{J})$ for $\mathcal{J} \subset \mathcal{I}$. Consequently, the candidate generation and pruning techniques of the standard Apriori framework can be used to find all frequent itsemsets, i.e., all itemsets whose support exceeds a user-defined threshold [1].

To compute the support of an itemset, we adopt some ideas that were presented in [11] for the binary case and can easily be extended to the fuzzy case.

Suppose that, for a given itemset $\mathcal{I}$, the concordance degrees $\tilde{C}(\boldsymbol{u}_i, \boldsymbol{u}_j)$ are stored in an $|N| \times |N|$ matrix. From this matrix, $\text{supp}(\mathcal{I})$ can easily be computed by summing all entries. Moreover, given the matrices for two itemsets $\mathcal{I}$ and $\mathcal{J}$, the matrix for the union $\mathcal{I} \cup \mathcal{J}$ is obtained by a simple position-wise t-norm combination. This approach is appealing for programming languages specifically tailored to matrix computations. In general, however, the storage requirements will be too high, especially noting that the matrices are normally quite sparse. More efficient implementations should hence exploit dedicated techniques for handling sparse matrices that, amongst others, avoid the storage of zero entries.

For each itemset $\mathcal{I}$ exceeding the given support threshold, a set of candidate rules $\mathcal{A} \rightharpoonup \mathcal{B}$ is derived by splitting $\mathcal{I}$ into antecedent part $\mathcal{A}$ and consequent part $\mathcal{B}$. For reasons of comprehensibility, we restrict ourselves to the case $|\mathcal{B}| = 1$, i.e., to consequents with a single attribute. A candidate rule of that kind is presented to the user if it exceeds the confidence threshold. While the concordance of the rule, $\tilde{C}$, is already known, this decision requires the additional computation of the discordance $\tilde{D}$.

4.3 Illustration

To illustrate our method (a thorough empirical evaluation is precluded due to space restrictions), we applied it to the Wine Quality data set from the UCI repository, in which each data record corresponds to a red wine described in terms of 11 numerical attributes and a quality degree between 0 and 10. Each attribute was replaced by two fuzzy attributes `small` and `large` with membership degrees 1 (0) and 0 (1) for the smallest and largest value, respectively, and linearly interpolating in-between. Using $r = 0.1$, we found the following rules exceeding a confidence threshold of 0.6:

- The more fixed acids and the more alcohol, the better the quality.
- The more volatile acids and sulfur dioxides, the lower the quality.
- The more volatile acids and the less alcohol, the lower the quality.
- The more sulfur dioxides and the less sulfates, the lower the quality.
- The more sulfur dioxides and the less alcohol, the lower the quality.
- The more sulfates and alcohol, the better the quality.

Roughly, one can observe that the amounts of volatile acids, sulfates and alcohol seem to have the strongest influence on the quality of the wine, with the former in a negative and the latter two in a positive manner. These results seem to agree quite nicely with oenological theory [6].

5 Concluding Remarks

We have presented a unified framework for mining fuzzy gradual dependencies, in which the strength of association between itemsets is measured in terms of a fuzzy rank correlation coefficient. As explained above, this framework generalizes previous proposals and allows for a seamless transition from a purely qualitative to a quantitative assessment.

An important aspect to be addressed in future work concerns more efficient algorithms and implementations for mining gradual dependencies. Due to the need to compare *pairs* of observations, the inherent problem complexity increases from linear to quadratic in the size of the data set. Thus, in order to guarantee scalability, efficient pruning techniques are needed that avoid unnecessary comparisons. Since the concordance relation in rank correlation is in direct correspondence to Pareto-dominance in preference modeling, it might be interesting to exploit algorithms that have recently been developed for the computation of so-called *skylines* (Pareto sets) of a database [13].

References

1. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD*, pages 207–216, Washington, D.C., 1993.
2. F. Berzal, J.C. Cubero, D. Sanchez, M.A. Vila, and J.M. Serrano. An alternative approach to discover gradual dependencies. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(5):559–570, 2007.
3. U. Bodenhofer. Representations and constructions of similarity-based fuzzy orderings. *Fuzzy Sets and Systems*, 137:113–136, 2003.
4. U. Bodenhofer and M. Demirci. Strict fuzzy orderings with a given context of similarity. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 16(2):147–178, 2008.
5. U. Bodenhofer and F. Klawonn. Robust rank correlation coefficients on the basis of fuzzy orderings: Initial steps. *Mathware & Soft Computing*, 15:5–20, 2008.
6. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.
7. D. Dubois, E. Hüllermeier, and H. Prade. A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery*, 13(2):167–192, 2006.
8. D. Dubois and H. Prade. Gradual inference rules in approximate reasoning. *Information Sciences*, 61(1,2):103–122, 1992.
9. L.A. Goodman and W.H. Kruskal. *Measures of Association for Cross Classifications*. Springer-Verlag, New York, 1979.
10. E. Hüllermeier. Association rules for expressing gradual dependencies. In *Proc. PKDD–02*, pages 200–211, Helsinki, Finland, 2002.
11. A. Laurent, M.J. Lesot, and M. Rifqi. GRAANK: Exploiting rank correlations for extracting gradual itemsets. In *Proc. FQAS–09*, pages 382–393, 2009.
12. C. Molina, J.M. Serrano, D. Sanchez, and M. Vila. Measuring variation strength in gradual dependencies. In *Proc. EUSFLAT–07*, pages 337–344, 2007.
13. D. Papadias, Y. Tao, G. Fu, and B. Seeger. Progressive skyline computation in database systems. *ACM Trans. Database Syst.*, 30(1):41–82, 2005.