

TITLE OF DISSERTATION: *Attribute-Oriented Fuzzy Induction: Data Mining Approach*

NAME OF STUDENT: *Rafal A. Angryk*

NAME OF DEPARTMENT: *Electrical Engineering and Computer Science Department*

CHAIRMAN OF DISSERTATION: *Dr. Frederick E. Petry*

MEMBERS OF PH.D. COMMITTEE: *Dr. Bill P. Buckles, Dr. Marin Simina, Dr. Maria Cobb*

1. Introduction

The research area I would like to focus on is known as Attribute-Oriented Induction (AOI) for Knowledge Discovery in Databases. Attribute-Oriented Induction is a process of grouping of data, enabling transformation of similar data collections, expressed originally in a database at the low (primitive) level, into more abstract conceptual representations. Process of data generalization is a fundamental element of Attribute-Oriented Induction, a descriptive database mining technique [1], which compresses the original set of data into a generalized relation, providing concise and summarative information about the massive set of the low-level task-relevant data.

Generalization of database records is performed on an attribute-by-attribute basis, applying a separate concept hierarchy for the each of the generalized attributes included in the relation of task-relevant data. Each concept hierarchy represents background knowledge about the domain allowing gradual, similarity-based, aggregation of attribute values stored in the original tuples. This hierarchy is built usually in the bottom-up manner progressively increasing the abstraction of the generalization concepts at each new level. Original attribute values from the task-specific (initial) relation are gradually generalized to more abstract concepts. Such terse representation of dataset is then directly presented to the user or used as a input table for additional Data Mining techniques.

2. Summary of earlier work on the problem

The idea of applying concept hierarchies for attribute-oriented induction in data mining was popularized by Han and his co-researchers [2, 3, 4, 5] and extended further by Hilderman, Hamilton, Cecrone and their co-workers [6, 7, 8].

2.1. Crisp Concept Hierarchies [2-8]

Originally, hierarchical grouping proposed by Han [3] was based on tree-like generalization hierarchies, where each of the concepts at the lower level of the generalization hierarchy was allowed to have just one abstract concept at the level directly above it (its direct generalization). There was no consideration of the degree of this relationship.

2.2. Fuzzy Concept Hierarchies [13-17]

Yager [9, 10], followed by Kacprzyk [11] and Dubois and Prade [12], investigated application of fuzzy sets to the area of dataset summarization through linguistic concepts. Summarization of database records with utilization of fuzzy concept hierarchies was approached directly in late nineties by four groups of independent researchers. Lee and Kim [13] used ISA hierarchies, from area of data modeling, to generalize database records to more abstract concepts. Lee [14] applied fuzzy generalization hierarchies to mine generalized fuzzy quantitative association rules. Cubero, Medina, Pons and Vila [15] presented fuzzy gradual rules for data summarization and Raschia, Ughetto and Mouaddib [16, 17] implemented SaintEtiq system for data summarization through extended concept hierarchies.

Fuzzy hierarchy of concepts [13-17] reflects the degree with which one concept belongs to its direct abstract and more than one direct abstract of a single concept is allowed. Because of the lack of guarantee of exact vote propagation, such a hierarchy seems to be more appropriate for simplified data summarization, or to the cases when subjective results are to be emphasized (we purposely want to modify the role or influence of certain records).

2.3. Fuzzy Relational Databases

The similarity-based fuzzy model of a relational database, proposed first by Buckles and Petry in 1982 [18, 22-26]], is actually a formal generalization of the ordinary relational database model introduced by Codd [19] in 1970. The model, based on the max-min composition of a similarity relation utilized as the extension of the classical identity relation coming from the theory of crisp sets, was further extended by Shenoj and Melton in 1989 [20, 28-32] with the concept of the proximity relation.

The most distinctive qualities of the fuzzy relational database are: (1) allowing non-atomic domain values, when characterizing particular attributes of a single entity and (2) generation of equivalence classes (effecting such basic properties of relational database as the removal of redundant tuples) with the support of similarity relation [21] applied in the place of traditional identity relation.

3. Plan for the investigation

My objective is to investigate the development and properties of Crisp (CCH) and Fuzzy (FCH) Concept Hierarchies and their potential in attribute-oriented generalization of ordinary and fuzzy relational databases. The following chapter includes a detailed explanation of this idea.

There are three aspects of my dissertation research. As I mentioned earlier there was some work done on applying FCH in area of ordinary relational databases, I want extend these concepts, adding the features, which I believe, are important for consistent DM via AOI with FCH. Main part of my work is to apply the AOI approach to mine knowledge from fuzzy databases, which has not appeared previously in the literature.

3.1. Fuzzy Concept Hierarchies in Ordinary Databases - Research Task

The problem appears to lay in the lack of fuzzy model completeness, where a single tuple is guaranteed to remain with a “weight” of one record at each of the fuzzy generalization levels. When generalizing data for data mining purposes we have to preserve the number of tuples and the relation between them in the identical proportions at each level of generalization. This leads to the two following properties, which must be maintained at each level of generalization hierarchy: (1) the set of concepts at each level of hierarchy should cover all of the attribute values that occurred in the original database (so we are guaranteed to not lose the number of tuples when generalizing their values), (2) never allow any attribute value (or its abstract) to be counted more or less than once at each level of the generalization hierarchy (when we allow a concept to partially belong to more than one of its direct abstracts, we have to check each time that the sum of fractional memberships is equal 1.0). This conclusion derives directly from the Vote Propagation principle introduced by Han [3]

Expected solution: The Fuzzy Concept Hierarchy is *complete and consistent* (preserves Exact Vote Propagation principle) when for all adjacent hierarchy levels, S and T (where T is a direct abstraction level of S), the following relationship is satisfied: $\sum_{i=1}^{|T|} m_{st_i} = 1, \forall s \in S, \forall t_i \in T$

where: i is the index of the generalization concepts at the abstraction level T .

In other words, the sum of weights assigned to the links leaving a single node in the fuzzy concept hierarchy needs to be always 1.0 to preserve completeness of the generalization model.

3.2. Crisp Concept Hierarchies in Fuzzy Databases - Research Task

I am interested in employing the similarity/proximity tables, which are essential elements in each of the mentioned fuzzy database models, as a source of information sufficient to form the crisp concept hierarchies for attribute-oriented generalization purposes. Consequently, this approach should allow us to perform attribute-oriented induction without the necessity of obtaining the background knowledge about each of generalized attributes, presumed so far to be obligatory in this method.

Expected solution: The existence of a similarity relation modeled for a particular domain can lead to the extraction of a crisp concept hierarchy, allowing attribute-oriented generalization. Let S_a be the a -cut of the similarity relation S , presented in the Table 1. It can be shown [21] that if S is a similarity relation on a given domain D_j (which is a single attribute in our case), then $\forall a \in (0,1]$ each S_a creates equivalence classes in the domain D_j . Now, let P_a denote the equivalence class partition induced on domain D_j by S_a . Clearly, $P_{a'}$ is a refinement of P_a if $a' \geq a$. A nested sequence of partitions $P_{a^1}, P_{a^2}, \dots, P_{a^k}$ may be represented diagrammatically in the form of a *partition tree*.

The nested sequence of partitions in the form of a tree has a structure identical with the crisp concept hierarchy for data mining generalization purposes.

3.3. Fuzzy Concept Hierarchies in Ordinary Databases - Research Task

An acceptance of non-atomic values may lead to the occurrence of imprecision, taking place when single entity is described by multiple values, which are not considered to be equal at the specific level of similarity/proximity. I am interested in developing a method allowing attribute-oriented generalization of multiple attribute values, describing a single entity, according to their similarity. There is a connection between precision and certainty since very imprecise (abstract) statements have a greater chance to be correct than precise ones [37, 38], I would like to investigate if AOI applied on imprecise data would lead to removing the imprecision when the specific level of abstraction in the concept hierarchy is reached during generalization.

Expected solution: I believe that during generalization of tuples with imprecision it would be preferable to take two basic things under consideration:

- a) *Value of similarity between the terms (attribute values) describing particular entity.* As I mentioned before a strong connection between precision and certainty seems to exist.
- b) *Number of similar terms,* since it can influence the certainty about the generalization path.

Driven by these two assumptions I believe that generalization of the imprecise information should be assessed both on the number of inserted descriptors for particular domain as well as on the similarity of inserted values. This is just a sketch of possible solution; a mechanism performing such generalization of multiple attribute-values still needs to be investigated.

References

- [1] A. Feelders, H. Daniels & M. Holsheimer, Methodological and practical aspects of data mining, *Information & Management*, 37, 2000, 271-281.
- [2] J. Han & M. Kamber, *Data Mining: Concepts and Techniques*, New York: Morgan Kaufmann, 2000.
- [3] J. Han, Y. Cai, & N. Cercone, Knowledge discovery in databases: An attribute-oriented approach, *Proc. 18th Int. Conf. Very Large Data Bases*, Vancouver, Canada, 1992, 547-559.
- [4] J. Han & Y. Fu, Exploration of the Power of Attribute-Oriented Induction in Data Mining, in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (Menlo Park, CA: AAAI/MIT Press, 1996) 399-421.
- [5] J. Han, Mining Knowledge at Multiple Concept Levels, *Proc. 4th Int'l Conf. on Information and Knowledge Management*, Baltimore, Maryland, 1995, 19-24.
- [6] H.J. Hamilton, R.J. Hilderman & N. Cercone, Attribute-oriented induction using domain generalization graphs, *Proc. 8th IEEE Int'l Conf. on Tools with Artificial Intelligence*, Toulouse, France, 1996, 246-253.
- [7] C. L. Carter & H. J. Hamilton, Efficient Attribute-Oriented Generalization for Knowledge Discovery from Large Databases, *IEEE Transactions on Knowledge and Data Engineering*, 10(2), 1998, 193-208.
- [8] R. J. Hilderman, H.J. Hamilton & N. Cercone, Data mining in large databases using domain generalization graphs, *Journal of Intelligent Information Systems*, 13(3), 1999, 195-234.
- [9] R. R. Yager, On linguistic summaries of data, in W. Frawley & G. Piatetsky-Shapiro (Eds.), *Knowledge Discovery in Databases* (Menlo Park, CA: AAAI/MIT Press, 1991) 347-363.
- [10] D. Rasmussen & R.R. Yager, A Fuzzy SQL Summary Language for Data Discovery, in D. Dubois, H. Prade and R.R. Yager (Eds.), *Fuzzy Information Engineering: A Guided Tour of Applications* (New York: John Wiley and Sons Inc., 1997) 253-264.
- [11] J. Kacprzyk, Fuzzy logic for linguistic summarization of databases, *Proc. 8th Int'l Conf. on Fuzzy Systems*, Seoul, Korea, 1999, 813-818.
- [12] D. Dubois & H. Prade, Fuzzy sets in data summaries - outline of a new approach, *Proc. 8th Int'l Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Madrid, 2000, 1035-1040.
- [13] D. H. Lee and M.H. Kim, Database summarization using fuzzy ISA hierarchies, *IEEE Transactions On Systems, Man, and Cybernetics - part B*, 27(1), 1997, 68-78.
- [14] K.-M. Lee, Mining generalized fuzzy quantitative association rules with fuzzy generalization hierarchies, *Proc. Joint 9th IFSA World Congress and 20th NAFIPS Int'l Conf.*, Vancouver, BC, Canada, 2001, 2977 -2982.
- [15] J. C. Cubero, J.M. Medina, O. Pons, M.A. Vila, Data Summarization in Relational Databases Through Fuzzy Dependencies, *Information Sciences*, 121(3-4), 1999, 233-270.
- [16] G. Raschia, L. Ughetto, N. Mouaddib, Data summarization using extended concept hierarchies, *Proc. Joint 9th IFSA World Congress and 20th NAFIPS Int'l Conf.*, Vancouver, BC, Canada, 2001, 2289 -2294.
- [17] G. Raschia, N. Mouaddib, *SAINTETIQ: a fuzzy set-based approach to database summarization*, *Fuzzy Sets and Systems* 129 (2002) 137 -162.
- [18] B.P. Buckles, F.E. Petry, A fuzzy representation of data for relational databases, *Fuzzy Sets and Systems*, 7, 1982, 213-226.
- [19] F.E. Codd, A relational model of data for large share data banks, *Comm. ACM* 13 (1970) 377{387}.

- [20] S. Sheno and A. Melton, Proximity Relations in the Fuzzy Relational Database Model, *International Journal of Fuzzy Sets and Systems*, 31 (3), 1989, 285-296.
- [21] L.A. Zadeh, Similarity relations and fuzzy orderings, *Information Sciences*, 3, 1970, 177-200.
- [22] B.P. Buckles, F.E. Petry, Fuzzy databases and their applications. In: M.M. Gupta and E. Sanchez, Editors, *Fuzzy Information and Decision Processes*, North-Holland, Amsterdam, Abstract-Compendex (1982), 361–371.
- [23] B.P. Buckles, F.E. Petry, Information-theoretic characterization of fuzzy relational databases, *IEEE Transactions on Systems, Man and Cybernetics*, 13, 1983, 74–77.
- [24] B.P. Buckles, F.E. Petry, Extending the fuzzy database with fuzzy numbers. *Information Science* 34, 1984, 145–155.
- [25] B.P. Buckles, F.E. Petry and H.S. Sachar, A domain calculus for fuzzy relational databases. In: (25th ed.), *Technical Report CSE TR-85-006*, Department of Computer Science and Engineering, University of Texas at Arlington (1985).
- [26] F.E. Petry, *Fuzzy Databases: Principles and Applications*, Kluwer Academic Publishers, 1996, Boston.
- [27] J.C. Bezdek and J.D. Harris, Fuzzy Partitions and Relations: an Axiomatic Basis for Clustering, *Fuzzy Sets and Systems*, vol.1, pp. 111-127, 1978.
- [28] A. Melton and S. Sheno, Fuzzy Relations and Fuzzy Relational Databases, *International Journal of Computers and Mathematics with Applications*, 21 (11/12), 1991, 129-138.
- [29] S. Sheno and A. Melton, An Extended Version of the Fuzzy Relational Database Model, *Information Sciences*, 52 (1), 1990, 35-52.
- [30] S. Sheno, A. Melton, and L. T. Fan, An Equivalent-Class Model of Fuzzy Relational Databases, *The International Journal of Fuzzy Sets and Systems*, 38, 1990, 153-170.
- [31] S. Sheno, A. Melton, and L. T. Fan, Functional Dependencies and Normal Forms in the Fuzzy Relational Database Model, *Information Sciences*, 60, 1992, 1-28.
- [32] S. Kumar De, R. Biswas, A. R. Roy, On extended fuzzy relational database model with proximity relations, *Fuzzy Sets and Systems*, 117, 195-201, 2001.
- [33] R.A. Angryk, F.E. Petry, Consistent fuzzy concept hierarchies for attribute generalization, *submitted to IASTED International Conference on Information and Knowledge Sharing (IKS 2003)*, 2003.
- [34] R.A. Angryk, F.E. Petry, *Data Mining Fuzzy Databases Using Attribute-Oriented Generalization*, *submitted to 2003 IEEE International Conference on Data Mining (ICDM 2003)*, 2003.
- [35] D. Dubois, H. Prade, *Fuzzy Sets and Systems: Theory and Applications*, Academic Press, New York, 1980.
- [36] D. Dubois, H. Prade, J.-P. Rossazza, Vagueness, typicality and uncertainty in class hierarchies, *International Journal of Intelligent Systems*, Vol. 6, 1991, 167-183.
- [37] S. Parsons, Current approaches to handling imperfect information in data and knowledge bases, *IEEE Transactions on Knowledge and Data Engineering*, 8(3), 1996, 353-372.
- [38] A. T. Berztiss, Uncertainty Management, in S K Chang (Ed.) *Handbook of Software Engineering & Knowledge Engineering*, Vol. 2, ISBN: 981-02-4974-8, 2002.