ICEB 2004 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-5-2004

Discovering Fuzzy Functional Dependencies as Semantic Knowledge in Large Databases

Xue Wang

Guoqing Chen

Discovering Fuzzy Functional Dependencies as Semantic Knowledge in Large Databases

Xue Wang , Guoqing Chen

School of Economics and Management, Tsinghua University, Beijing 100084, China
{wangx3.03, chengq}@em.tsinghua.edu.cn

ABSTRACT

Fuzzy functional dependency (FFD) is a kind of semantic knowledge and can be discovered from a large volume of business data. Sectional FFD and Attribute FFD are discussed so as to reflect semantics of the business world and express useful information that is natural for people to comprehend. The experimental results on an insurance data set show that the proposed method can extract knowledge efficiently and effectively.

Keywords: data mining, knowledge discovery, fuzzy functional dependency, Sectional FFD, Attribute FFD.

1. INTRODUCTION

Widespread internet applications and e-business practices have resulted in a rapid growth of the database size that is beyond the scope of expert human capabilities to scan all the collected data and to discover the useful knowledge hidden in it. Therefore, data mining and knowledge discovery, for finding interesting patterns, dependencies, summaries, regularities, etc, has become an increasing important subject in the research area of database and business intelligence.

Functional dependency (FD) is important in both database design and maintenance. The classical definition of FD is: X functionally determines Y, (or Y is functionally dependent on X), denoted by X→Y, if and only if $\forall t_1,t_2 \in$ R, if $t_1(X) = t_2(X)$ then $t_1(Y) = t_2(Y)$. But in real world applications, information is often incomplete or ambiguous. For example, customers may not be willing to provide their actual age but "about twenty", "middle aged", "25-30" or the like. With the inception of fuzzy logic [13] to model imprecise information, fuzzy extensions has been introduced into functional dependency in different aspects since 1980's (Prade et al. [5] Raju et al.[6] Bhuinya et al.[1], Chen et al.[2][3], Cubero et al[4], Saxena et al.[7], S.Ben Yahia et al.[11], etc). A general setting of fuzzy functional dependencies proposed by Chen et al is as follows [2]:

Let U be the set of all attributes for a relation scheme R and X,Y $\subseteq$ U. X functionally determines Y (or Y is functionally dependent on X) to the degree ϕ, denoted by $X \rightarrow_{\phi} Y$, if and only if for any tuples t_1,t_2,

$$\min_{t_1,t_2\in R} I(c(X(t_1),X(t_2)),c(Y(t_1),Y(t_2))) \geq \phi \qquad (1)$$

where $\phi \in [0,1]$. ***I*** is a fuzzy implication operator (FIO) and c is an equality measure.

In a similar spirit, our previous work [12] presented a specific type of fuzzy functional dependency based on tuples and label closeness measures, which has a good arithmetic efficiency in the context of data mining. In addition, it can reflect both overall and partial knowledge of the data. For example, we may get a FFD at the attribute level (Attribute FFD), such as Age ~> Salary (Age fuzzy determines Salary), but some times, we may only get FFD for sub-classes of attribute values (Sectional FFD), such as Age(young) ~> Salary(low). Hence the construction and discovery of Sectional FFD is useful and novel and can serve as a type of interesting pattern and aid to enrich the knowledge base of a company.

2. SECTIONAL FFDs AND ATTRIBUTE FFDs

In this section, we discuss the notions and properties of Sectional FFDs and Attribute FFDs, viewed as a kind of semantic knowledge.

2.1 Notions

When dealing with fuzzy data, linguistic labels are defined in a unified way with membership grades. Examples of these labels are small, large and young. In some situations, such labels are used to represent abstract and linguistic summarization of quantitative data values. Thus, the data concerned can be represented in the form of a vector like $V(\mu_1/L_1,\ \mu_2/L_2\ \ldots\ \mu_n/L_n)$, where $L_1 \ldots L_n$ are labels and $\mu_1 \ldots \mu_n$ are membership grades. Next, we replace this vector with the maximal membership grade and its corresponding label μ_i/L_i where $\mu_i=\max(\mu_1 \ldots \mu_n)$. For instance, "about 50" can be translated into "0.9/old".

Definition 1 Let U be the set of all attributes for a relation scheme R, and X,Y $\in$ U, $L_1 \ldots L_n$ are labels of X and $L'_1 \ldots L'_n$ are labels of Y. For a tuple t in R, its X and Y values are μ_i/L_i and μ'_j/L'_j. θ, $\alpha \in [0,1]$, where α is a given threshold. t satisfies a tuple relation to the degree θ, denoted by $X(t, L_i) \leadsto_{\theta} Y(t, L'_j)$ if and only if

$$\theta = \boldsymbol{I}\,(\mu_i, \mu_j') \geq \alpha \qquad (2)$$

where *I* is a fuzzy implication operator (FIO). Most used FIOs are Lukasiewicz operator, Kleene-Dienes operator, Gödel operator, R_0 operator, etc. Here, we choose R_0

implication operator [8] as an example to derive a specific form of tuple relation which has some good properties as stated in section 2.3.

$$I_{R_0}(a,b) = \begin{cases} 1 & \text{if } a \le b \\ \max(1\text{-}a,b) & \text{otherwise} \end{cases} \quad (3)$$

Definition 2 Label closeness measure which describes the relationship between two given labels denoted by $\sigma(L_i,L_j)$, satisfies the following properties:

1) $\sigma(L_i,L_i) = 1$ (4)
2) $\sigma(L_i,L_j) = \sigma(L_j, L_i)$ (5)

More specifically, for $L^*=\{L_1\ldots L_n\}$, let $\Theta(L^*) = \min(\sigma(L_i,L_j) \mid L_i,L_j \in L^*)$ and for $\ell^{**}= \{L_1^*, L_1^*\ldots L_n^*\}$, which is a set of sets, $\Theta(\ell^{**}) = \min(\sigma(L_1^*), \sigma(L_1^*) \ldots \sigma(L_n^*))$.

Definition 3 For any label L_i of X, all tuples satisfying the tuple relation $X(t, L_i) \sim>_\theta Y(t, L'_j)$ compose a set T_i, more formally, $T_i=\{t \mid X(t, L_i) \sim>_{\theta i} Y(t, L'_j)\}$. All those L'_j compose a set named L_i^*. A Sectional FFD, denoted by $X(L_i) \sim>_\theta Y(L_i^*)$ is valid if and only if

$$\Theta(L_i^*) \ge \beta \quad (6)$$

where β is a given threshold between 0 and 1, $\theta = \min(\theta_i)$.

Definition 4 An Attribute FFD (AFFD) $X \sim>_{\theta'} Y$ holds if and only if for any two label L_i and L_j, the Sectional FFDs $X(L_{i/j}) \sim>_\theta Y(L_{i/j}^*)$ hold and

$$I'(\sigma(L_i,L_j), \sigma(L_i^*,L_j^*)) \ge \omega \quad (7)$$

where I' is a FIO and ω is a given threshold. ($\theta' = \min(\theta)$).

Here we also use R_0 to derive a specific form of FFD. We can easily see, if L* is a single element set, the Sectional FFD $X.L_i \sim> Y.L^*$ mentioned above is surely valid.

2.2 An Example

The following table is part of a fuzzy relational database. The membership functions of attribute labels are given in figure 2.(with ω=0.7, α = 0.7)

T#	Age	Salary
t1	Young	1200
t2	35	0.8/low
t3	About 30	About 1500
t4	50-55	3800
t5	0.9/old	5000
t6	60	0.9/high

Table 1 Example data table

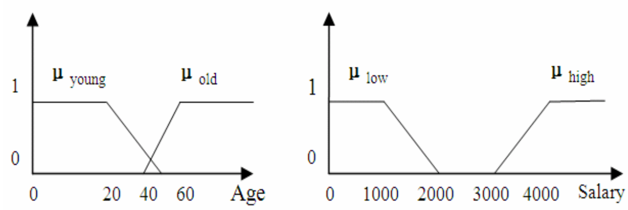

Figure 2 Membership functions

The first step: transform all values in table 1 into table 2 using maximal membership grade and its corresponding label.

T#	Age	Salary
t1	1/young	0.8/low
t2	0.75/young	0.8/low
t3	0.7/young	0.72/low
t4	0.75/old	0.8/high
T5	0.9/old	1/high
T6	1/old	0.9/high

Table 2 Transformed data

The second step: check if for every tuple, the tuple relation holds. We can easily derive that Age(t_1,young) $\sim>_{0.8}$ Salary(t_1, low) , Age(t_2,young) $\sim>_1$ Salary(t_2,low), Age(t_3,young) $\sim>_{0.72}$ Salary(t_3,low), Age(t_4,old) $\sim>_1$ Salary(t_4,high), Age(t_5,old) $\sim>_1$ Salary(t_5,high), Age(t_6,old) $\sim>_{0.9}$ Salary(t_6,high). If the threshold α = 0.7 then all tuple relations hold.

The third step: for attribute Age, label young has three tuples t_1,t_2 and t_3, the corresponding label in attribute Salary L*={low}, L* is a single element set, so the sectional FFD Age(young)$\sim>_{0.72}$ Salary(low) holds. In the same way, we can easily derive Age(old)$\sim>_{0.9}$ Salary(high), in addition, $I_{R0}(\sigma(\text{young,old}), \sigma(\text{low, high}))$ $=1>\omega$, then the Attribute FFD Age $\sim>_{0.72}$ Salary holds.

Notably, if the salary of the first tuple is 0.8/high instead of 0.8/low, we will get Age(t_1,young)$\sim>_{0.8}$ Salary(t_1,high). Therefore, for the left side label "young", L*={low,high}, we have to compute σ(low,high). If σ(low,high) is below the given threshold, the Sectional FFD Age(young) ~> Salary(low, high) is denied.

2.3 Properties

It can be proved that Attribute fuzzy functional dependencies with R_0 implication operator have the following important properties.

1) **Reflexivity:** if $Y \subseteq X \subseteq U$, where X is a set of attributes, then X ~>Y holds. It is a trivial fuzzy functional dependency.
2) **Augmentation:** if X~>Y holds, then for $Z \subseteq U$, XZ~>YZ holds.
3) **Union Rule:** if X~>Y, X~>Z, then X~>YZ.
4) **Decomposition Rule:** if X~>Y , $Z \subseteq Y$, where Y is a set of attributes, then X~>Z.
5) **Partial Transitivity:** if $X \sim>_\tau Y$ and $Y \sim>_\eta Z$ with $\tau, \eta \ge 1/2$, then $X \sim>_\gamma Z$ holds with $\gamma=\min(\tau, \eta)$.

3. MINING ALGORITHM

In this section, we'll present a mining algorithm to discover Sectional FFDs and Attribute FFDs.

As shown in section 2.3, the AFFD X~>YZ can be decomposed in to X~>Y and X~>Z. Therefore we will deliberate to find non-trivial AFFDs, each with a single attribute in its right-hand side. We can derive the minimal FFD set using a map listed below. For example, the line between X and XY means we will check whether the AFFD X~>Y holds, in the same way, the line between XY and XYZ means we will check whether the AFFD XY~>Z holds. An algorithm of the mining procedure for X~>Y is as follows.

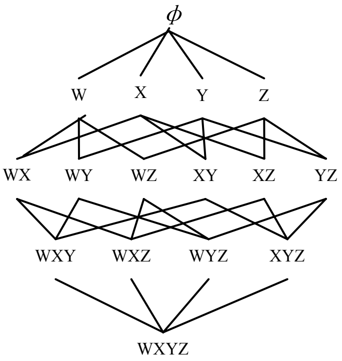

Figure 3 Map for mining

INITIALIZING:
for each Label L_k of X
    L_k *:= Φ ;
    sectional_flag(k) := valid;
    attribute_flag := valid;
end for

MINING
for each t in R
    get corresponding X,Y value μ_i/L_i , μ'_j/L'_j;
    compute $\theta := \boldsymbol{I}(\mu_i, \mu_j')$;
    if $\theta < \alpha$
        tuple relation is denied;
        sectional_flag(i) := denied;
    else
        add L_j' into the set L_i*;
    end if
end for
for every label L_i of X
    if sectional_flag(i) = valid & Θ (L_i*) $\geq \beta$
        the sectional FFD X(L_i)~>Y(L_i*) holds;
    else
        attribute_flag := denied;
    end if
end for
if attribute_flag = valid
    for every two labels L_i,L_j of X
        if $I'(\sigma(L_i,L_j), \sigma(L_i^*,L_j^*)) < \omega$
            attribute_flag := denied;
            break;
        end if
    end for
end if
if attribute_flag = valid
    the Attribute FFD X~>Y holds
end if

When mining such FFDs from databases, we need to check every tuple to see if it satisfies tuple relation and compute label closeness value. For X~>Y, if X has M_1 labels and Y has M_2 labels, the algorithm's complexity is at $O(N+M_1^2+M_2^2)$. In general, M_1, M_1<<N, so valuable patterns could be discovered quite efficiently.

4. EXPERIMENTAL RESULTS

We have carried out an experiment on a real business data set http://www.smr.nl. The algorithm was developed in C language, and run on PC with PIII866, RAM 256M, and Windows 2000 professional.

4.1 Data Description

The data set is a real set of an insurance company, provided by "Dutch Data Mining company sentient machine research". It contains 5822 tuples (transactions) and 86 attributes such as "number of houses", "customer main type", "average size household", etc. Because of space limitation, we only present the results discovered between the first 10 attributes within all tuples. Table 3 gives a general description of these first 10 attributes.

No	Description	Domain
1	Customer Subtype	1 High Income, expensive child … 15 Household with children … 41 Mixed rurals
2	Number of houses	Number of 1-10
3	Avg size household	1-6 (1 is the smallest, 6 the largest)
4	Avg age	1 20-30 years … 6 70-80 years
5	Customer main type	1 Successful hedonists … 8 Family with grown ups … 10 Farmers
6	Roman catholic	0 0% 1 1 - 10% … 9 100%
7	Protestant	0 0% … 9 100%
8	Other religion	0 0% … 9 100%
9	No religion	0 0%... … 9 100%
10	Married	0 0% … 9 100%

Table 3 Data Description

4.2 Data Preparation

Among the descriptions above, we can see that there are fuzzy values, such as "Avg size household", graded values, such as "Avg age" and crisp values, such as "Customer Subtype". We first transformed all fuzzy terms or graded figures into a unified type – linguistic labels with corresponding membership grades. For example, there are six ranks in the attribute of "Avg size household" – "1, the smallest and 6 the largest". We transfer them into 3 labels as "small, middle, large", and the original graded values can be translated as shown in Table 4.

Rank	Label/value
1	small/1
2	small/0.9
3	medium/0.9
4	medium/1
5	large/0.9
6	large/1

Table 4 Fuzzy values of "Avg size household"

At the same time, we defined label closeness measures. For example σ (small, medium) = 0.8, σ (medium, large) = 0.8, σ (small, large) = 0.

4.3 Results

For illustrative purposes, we only list several sectional FFDs (Table 5) and Attribute FFDs (Table 6) with top θ values.

No	Sectional FFDs
1	Sub type [15] $\sim>_1$ number of houses [many]
2	Sub type [15] $\sim>_1$ age [old]
3	Sub type [15] $\sim>_1$ protestant [medium]
4	age [young] $\sim>_1$ number of houses [small]
5	Number of houses[huge]$\sim>_1$size of household [small]
6	Protestant [large]$\sim>_{0.9}$roman catholic [very small, small]

Table 5. Sectional FFDs

No	Sectional FFDs
1	sub type $\rightarrow_1$ main type
2	number of houses $\rightarrow_1$ size of household
3	size of household $\rightarrow_1$ number of houses
4	size of household $\rightarrow_1$age
5	age $\rightarrow_1$ number of houses
6	age $\rightarrow_1$size of household
7	main type $\rightarrow_1$ number of houses
8	roman catholic $\rightarrow_{0.9}$ number of houses
9	married $\rightarrow_{0.9}$ number of houses

Table 6. Attribute FFDs

From the Attribute FFDs discovered, we can find some interesting patterns. For example, "number of houses" are determined by "size of household", "avg age", "main type", "roman catholic", "married", etc. Therefore "number of houses" is dependent on other attributes to a great extent. As to Sectional FFDs, they can reflect dependencies that hold partially. We can find that among Sectional FFDs, customers of "sub type 15" have many other behaviors in common. They are all old, have many houses, more than 50% of them are protestant. Such knowledge is deemed interesting and novel, which may provide decision-makers with a better understanding of the customers.

5. CONCLUSION

In this paper, we have discussed Attribute FFD and Sectional FFD as a kind of semantic knowledge which can reflect overall or partial knowledge of the data in a manner that is natural for people to comprehend. The experimental results on an insurance data set showed that the proposed method can extract knowledge efficiently and effectively.

6. ACKNOWLEDGEMENT

The work was partially supported by the National Science Foundation of China (70231010, 70321001) and MOE Funds for Doctoral Programs (20020003095).

REFERENCES

[1] B. Bhuniya, P. Niyogi, Losseless join property in fuzzy relational databases, *Data and Knowledge Engineering, 1993, pp. 109-124.*

[2] G.Q.Chen, J.Vandenbulcke, E.E.Kerre, A step towards the theory of fuzzy relational database design. *Proceedings on Computer, Management science. IFSA'91, pp 44-47.*

[3] G. Q. Chen, E. E. Kerre, J. Vandenbulcke, Normalization based on fuzzy functional dependency in a fuzzy relationnnal data model. Information Systems，Vol. 21, No.3, 299-310, 1996.

[4] J.C.Cubero, M.A.Vila, A new definition of fuzzy functional dependency in fuzzy relational databases. *International Journal of Intelligent Systems 9(5)(1994), 441-44.*

[5] H. Prade, C. Testemale, Generalizing database relational algebra for the treatment of incomplete or uncertain information and vague queries, *Information Sciences 34 (1984) 115-134.*

[6] K.S.V.N.Raju, A.K.Majumadar, Fuzzy functional dependencies and lossless join decomposition of fuzzy relational database systems, *ACM Transactions on Database Systems 13(2), 1988, 129-166*

[7] P.C. Saxena, B.K. Tyagi, Fuzzy functional dependencies and independencies in extended fuzzy relational databases models, *Fuzzy Sets and Systems 69 (1995) 65-89.*

[8] Guojun Wang, A formal deductive system for fuzzy propositional calculus. *Chinese science bulletin Vol.42 No.18, September 1997*

[9] S.L.Wang, J.W.shen, T.P.Hong, Mining fuzzy functional dependencies from quantitative data. *2000 IEEE*

[10] S.B.Yahia, A.Jaoua. Mining linguistic summaries of databases using based Lukasiewicz implication fuzzy functional dependency. *1999 IEEE*

[11] S.Ben Yahia, H.Ounalli, A.Jaoua, An extension of claical functional dependency: dynamic fuzzy functional dependency. 1999 Elsevier Science Inc.

[12] Xue Wang, A Simplified Fuzzy Functional Dependency and its Mining Method, International Conference on Fuzzy Information Processing Theories and Applications, March 1-4, 2003, Beijing.

[13] Zadeh, L. A. Fuzzy sets, *Information and control*, 1965, 8, 338-35