Functional dependencies in presence of errors

János Demetrovics
Comp. and Autom. Institute
Hungarian Academy of Science
Kende u. 13-17, H-1111, Hungary
dj@ilab.sztaki.hu

Gyula O.H. Katona and Dezső Miklós
Alfréd Rényi Institute of Mathematics
Hungarian Academy of Sciences
Budapest P.O.B. 127 H-1364 Hungary
ohkatona@renyi.hu, dezso@renyi.hu

January 26, 2012

Abstract

A relational database D is given with Ω as the set of attributes. The rows (tuples, data of one individual) are transmitted through a noisy channel. It is supposed that at most one data in a row can be changed by the transmission. We say that $A \to b$ ($A \subset \Omega, b \in \Omega$) is an error-correcting functional dependency if the data in A uniquely determine the data in b in spite of the error. We investigate the problem how much larger a minimal error-correcting functional dependency can be than the original one.

1 Introduction

Let us give some examples. Suppose that the pair of attributes (first name, last name) is a key in the database M. The values in M are the real data. However, some of the data can be erroneous: the information is misunderstood in a phone conversation, the typist makes a mistake or the informant simply lies. Say, if one of the first names "Mario" is replaced by "Maria" (M contains "Mario", M^* contains "Maria") then we might have two individual with names Maria Sklodowska in M^*, hence the individual (row) cannot be determined from these two attributes. The question raised here is what other additional attributes we need to make us able to determine the real person (row).

A database can be considered as an $m \times n$ matrix M, where the rows are the data of one individual, the data of the same sort (*attributes*) are in the same column. Denote the set of attributes (equivalently, the set of columns of the matrix) by Ω, its size is $|\Omega| = n$. It will be supposed that the data of two distinct individuals are different, that is, the rows of the matrix are different. Let $A, B \subset \Omega$. We say that B *funtionally depends* on A and write $A \to B$ if any two rows coinciding in the columns of A are also equal in the columns of B. Specially, if $K \to \Omega$ then K is called a *key*. In other words, there are no two distinct rows of the matrix which are equal in K. A key is a *minimal key* if no proper subset of it is a key. Denote the family of all minimal keys by $\mathcal{K}$.

Let M denote the matrix of the real data. These data are transmitted through a noisy channel. M^* ($m \times n$, again) denotes the matrix of the data obtained after the transmission. We know that M and M^* differ in at most e entries in each row. Although it is also supposed here that the real data of two distinct individuals are different, that is the rows of M are different, this cannot be stated about M^*.

We assume that the structure of M is known, on the other hand we only know the received rows of M^* and our aim is to make conclusions based on this information. Suppose for instance that $A \to a$ ($A \subset \Omega, a \in \Omega$) holds in M. Then the data in a row in the columns of A determine the data of the same row in the column a. We know however only the corresponding rows in M^*. The data in the columns of A do not necessarily determine the data in a, since these data are distorted. Can we enlarge A into an A' whose data (in M^*) already determine a? If yes, to what extent should it be enlarged?

For instance, if the number of errors in one row is at most one ($e = 1$) and sex is one of the attributes then either the first name or the sex is correct. Yet, (Maria, Sklodowska, M) can be found in two different rows of M^* (one was (Mario, Sklodowska, M) the other one was (Maria, Sklodowska, F)). Further attributes migh be needed to identify the individual. That is, (first name, last name, sex) is not a 1-error-correcting key.

Let us emphasize that the problem of data mining in case of errors is different. Then only M^* is known, nothing is known about M. In our case the keys, functional dependencies are known for us in M. We want to modify them only for the purposes in M^*. In case of data mining we have no prior information on M, the keys or error-correcting keys should be determined only on the base of M^*.

Formalize our notions. We write $C \to \{e\}B$ if the values of the matrix M^* in the columns of C determine the values in the columns B uniquely, assuming that at most e errors occur in every row. (Actually, it is determined in M, the data in M^* can be erronous.) In other words, for any row r of M^* there are no two rows s and t of M both having Hamming distance less then or equal to e from r in the columns of C but being different in the columns of B. This is called an e-error-correcting functional dependency. The aim of the present paper is to find inequalities between the sizes of the sets occuring in the real functional dependencies and the e-error-correcting ones. Our previous paper [4] dealt with the case of the keys. The results of the present paper are analogous.

It is worth mentioning that $\{a\} \to \{1\}a$ does not hold, since the knowledge of the data in the column a does not give any information, it can be erronous. It does not determine the value in column a.

The number of different entries in two rows is called the *Hamming distance* of these two rows. The $m \times |C|$ submatrix of M determined by the set C of its columns is denoted by $M(C)$. Suppose that the Hamming distance of any two rows of $M(C)$, which are different in $a \in \Omega$, is at least $2e + 1$. Then the Hamming distance of any two rows of $M^*(C)$ is at least 1, that is, knowing the entries of the unreliable matrix in C determines the value in the column a, $C \to \{e\}a$ is an e-error-correcting functional dependency. Here we used the assumption that the functional dependencies of M are known. The converse is true, too: if the Hamming distance of two rows of $M(C)$, which are different in a, is at most $2e$ then it may happen that the rows are equal in $M^*(C)$, that is, $C \to \{e\}a$ is not true. We obtained the following

proposition.

Proposition 1.1 *$C \to \{e\}a$ ($C \subset \Omega, a \in \Omega$) is an e-error-correcting functional dependency iff the pairwise Hamming distance of the rows of $M(C)$, which are different in a, is at least $2e+1$.* □

2 Error-correcting functional dependencies

It is easy to see that if the pairwise Hamming distance of the rows of $M(C)$ being different in a is at least $2e$ then the knowledge of $M^*(C)$ detects the error (i.e. the presence of the error in M^*), but does not determine the data in a uniqely, i.e. there can be more then one rows of M having the same values in $M^*(C)$. This case is less interesting, but it makes worth introducing the more general definition: $C \to (d)a$ is called a *d-distance functional dependency* iff the pairwise Hamming distance of the rows of $M(C)$, which are different in a, is at least d.

The main aim of the present investigations is to find connections between the functional dependencies and the d-distance functional dependencies. The next proposition is the first step along this line. Let $\mathcal{F}_a$ be the family of minimal subsets F of Ω satisfying $F \to a$ (in M!).

Proposition 2.1 *$C \to (d)a$ ($C \subset \Omega, a \in \Omega$) is a d-distance functional dependency iff for any choice $a_1, \ldots, a_{d-1} \in C$ one can find an $F \in \mathcal{F}_a$ such that $F \subseteq C - \{a_1, \ldots, a_{d-1}\}$.*

Proof. The necessity will be proved in an indirect way. Suppose that there exist $a_1, \ldots, a_{d-1} \in C$ such that $C - \{a_1, \ldots, a_{d-1}\}$ contains no member of $\mathcal{F}_a$, that is, $C - \{a_1, \ldots, a_{d-1}\} \to a$ does not hold. Therefore there are two rows of M which are equal in $M(C - \{a_1, \ldots, a_{d-1}\})$ and are different in a. The Hamming distance of these two rows in $M(C)$ is less than d. The obvious contradiction completes this part of the proof.

To prove the sufficiency suppose, again in an indirect way, that $M(C)$ contains two rows with Hamming distance $< d$ and the rows are different in a. Delete those columns where these rows are different. We found a set $C - \{a_1, \ldots, a_{d-1}\}$ satisfying the condition that $M(C - \{a_1, \ldots, a_{d-1}\})$ contains two rows which are equal everywhere, but the rows are different in

a. Therefore $C - \{a_1, \ldots, a_{d-1}\} \to a$ is not true in M, $C - \{a_1, \ldots, a_{d-1}\}$ cannot contain a member of $\mathcal{F}_a$. □

The systems of functional dependencies were characterized in [1]. We prefer an equivalent description (see e.g. [3]) by the closure

$$\mathcal{L}(A) = \{a : a \in \Omega, A \to a\} \ (A \subseteq \Omega).$$

It is easy to see that this closure satisfies the following 3 conditions.

$$A \subseteq \mathcal{L}(A), \tag{i}$$

$$A \subseteq B \text{ implies } \mathcal{L}(A) \subseteq \mathcal{L}(B), \tag{ii}$$

$$\mathcal{L}(\mathcal{L}(A)) = \mathcal{L}(A). \tag{iii}$$

It is well-known ([1], [2]) that there is a database for any closure, in which the system of functional dependencies is exactly the one defined by this closure. This is why it is sufficient to give a closure rather than constructing the complete database or matrix.

It is possible to give a characterization with the families $\mathcal{F}_a$ as well. It is easy to see that $\mathcal{F}_a$ consists of $\{a\}$ and a (possibly empty) *inclusion-free* family of subsets of $\Omega - \{a\}$. (Inclusion-free means that $F_1, F_2 \in \mathcal{F}_a, F_1 \neq F_2$ implies $F_1 \not\subset F_2$.) We need one more condition for the interrelation between these families. Howvere, since we did not find the shortest form and no such characterization is needed in this paper we prove only the following lemma, which will be needed later.

Lemma 2.2 *Given an inclusion-free family $\mathcal{F}$ of subsets of a $\Omega \setminus \{a\}$, there is a system of functional dependencies (and therefore, with the preceeding remark, a relation M) such that it defines $\mathcal{F}_a = \mathcal{F} \cup \{\{a\}\}$ for some $a \in \Omega$.*

Proof. Fix an $a \in \Omega$ and define $\mathcal{F}_a$ as the family consisting of $\{a\}$ and $\mathcal{F}$ placed in some way in the remaing part of Ω. Let $\mathcal{L}(A) = A \cup \{a\}$ if $F \subseteq A$ for some $F \in \mathcal{F}_a$ and $\mathcal{L}(A) = A$ otherwise. It is easy to see that this function satisfies conditions (i)-(iii), that is, it is a closure. □

In other words, Lemma 2.2 says that for any inclusion-free family $\mathcal{F}$ on an $n-1$-element set there is a database where the family of minimal sets F satisfying $F \to a$ is exactly exactly equal to $\mathcal{F}_a = \mathcal{F} \cup \{\{a\}\}$.

Proposition 2.1 makes us able to give an abstract combinatorial definition, independent of databases. Let X be an n-element set and $\mathcal{F}$ be an inclusion-free family of its subsets. The *d-blownup* of $\mathcal{F}$ (in notation $\mathcal{F}(d)$) is defined by

$$\mathcal{F}(d) = \{G \subseteq X : \text{ for any choice of } x_1, \ldots, x_{d-1} \in G \ \ \exists F \in \mathcal{F} \text{ such that}$$

$$F \subseteq G - \{x_1, \ldots, x_{d-1}\} \text{ and } G \text{ is minimal for this property}\}.$$

Note that $\mathcal{F}(1) = \mathcal{F}$ and that, as we will see later, for an inclusion-free family of sets $\mathcal{F}$ forming the left hand sides of the functional dependencies of a relation $\mathcal{F}(d)$ will be the left hand sides of minimal d-distance dependencies. set

Our first observation is that it may happen that the d-blownup of $\mathcal{F}$ is an empty family while the original $\mathcal{F}$ is not. Fix an element $a \in X$ and an integer $2 \leq k$. Define $\mathcal{F}$ as the family of all k-element sets ($\subset X$) containing a. Then for any $C \subseteq X$ $C - \{a\}$ cannot contain any member of $\mathcal{F}$ therefore $\mathcal{F}(d)$ is empty for $2 \leq d$.

On the other hand, if $\mathcal{F}$ consists of all k-element subsets of X then all sets $G \subseteq X$ with $k + d - 1$ elements form $\mathcal{F}(d)$. Our last example suggests that the sizes of the members of $\mathcal{F}(d)$ do not exceed the sizes of the members of $\mathcal{F}$ by too much. We will show that this is not really true.

We say that the family $\mathcal{F}$ can be *pinned* by p elements if there are $x_1, \ldots, x_p$ such that no member of $\mathcal{F}$ avoids all of them, that is $F \cap \{x_1., \ldots, x_p\} \neq \emptyset$ $\forall F \in \mathcal{F}$. We saw that if $\mathcal{F}$ can be pinned by $d-1$ elements then $\mathcal{F}(d)$ is empty. Otherwise $\mathcal{F}(d)$ is never empty since X always satisfies the first part of the definition of the blownup and if it is not minimal, one can reduce it until arriving to a minimal set.

Theorem 2.3 *Let $n_0(k, d) \leq n$ and let $\mathcal{F}$ be an inclusion-free family of subsets of size at most k of a given set of size n, such that $\mathcal{F}$ cannot be pinned by $d-1$ elements. Then the sizes of the members of $\mathcal{F}(d)$ are at most $c_1 k^d$. On the other hand there is such an $\mathcal{F}$ for which all members of $\mathcal{F}(d)$ have size at least $c_2 k^d$. Here c_1 and c_2 depend only on d.*

As mentioned earlier, define now $\mathcal{F}_a(d)$ as the family of the left hand sides of the minimal d-distance dependencies described by Proposition 2.1. The family $\mathcal{F}(d)$ will then be defined as the union of the families $\mathcal{F}_a(d)$ for all

$a \in \Omega$. If $F \to a$ holds for some $(a \in \Omega, F \subseteq \Omega)$ and no proper subset F' of F satisfies $F' \to a$ then F is called a *minimal functional dependency set*. The following theorem will be obtained as an immediate consequence of the previous one.

Theorem 2.4 *Let $n_0(k, e) \leq n$. Suppose that all minimal functional dependency sets have sizes at most k. Then the members of $\mathcal{F}(2e+1)$ cannot be larger than $c_1 k^{2e+1}$. On the other hand there is a database with minimal functional dependency sets of size at most k in which all members of $\mathcal{F}(2e+1)$ have sizes at least $c_2 k^{2e+1}$. Here c_1 and c_2 depend on e, only.*

It is worth formulating the special case $e = 1$ with more specific constants.

Corollary 2.5 *Suppose that all minimal functional dependency sets have sizes at most k. Then the members of $\mathcal{F}(3)$ cannot be larger than $3k^3$. On the other hand there is a database with minimal functional dependency sets of size at most k in which all members of $\mathcal{F}(3)$ have sizes at least $c_2 k^3$ where c_2 is approximately $\frac{2}{27}$.*

Our conclusion is that the errors can considerably increase the sizes of the minimal fuctional dependencies, but the growth is only polynomial.

3 Proofs

Proof of Theorem 2.3 This proof is analogous to the proof of the main theorem of [4]. Let $\mathcal{F}$ be an inclusion-free family of subsets of X. The definition of $\mathcal{F}(d)$ implies that the family $\{F : F \in \mathcal{F}, F \subseteq G\}$ cannot be pinned by $d-1$ elements for members $G \in \mathcal{F}(d)$. On the other hand, by the minimality of a member $G \in \mathcal{F}(d)$, this is not true for $G - \{a\}$ where $a \in G$ is chosen arbitrarily. This gives the following proposition.

Proposition 3.1 *$G \in \mathcal{F}(d)$ iff $\{F : F \in \mathcal{F}, F \subseteq G\}$ cannot be pinned by $d-1$ elements, but $\{F : F \in \mathcal{F}, F \subseteq G - \{a\}\}$ can be pinned by some $d-1$ elements for every $a \in G$.*

□

Lower estimate. We give an inclusion-free family $\mathcal{F}$ consisting of $2 \le k$-element sets which generates an $\mathcal{F}(d)$ consisting of one member having size at least $c_2 k^d$.

Fix an integer $1 \le i < k$ and take a subset $A \subset X$ of size $i+d-1$. Let $B_1, B_2, \ldots$ be all the $\binom{i+d-1}{i}$ i-element subsets of A and

$$\mathcal{G}^i = \{B_1 \cup C_1, B_2 \cup C_2, \ldots\},$$

where $C_1, C_2, \ldots$ are disjoint subsets of $X - A$ with $|C_1| = |C_2| = \cdots = k - i$. This can be carried out if

$$i + d - 1 + \binom{i+d-1}{i}(k-i) \le n. \tag{3.1}$$

Using Proposition 3.1 we next show that the only member of $\mathcal{G}^i(d)$ is $D = A \cup \bigcup_i C_i$. It is easy to see that $\mathcal{G}^i$ cannot be pinned by $d-1$ elements.

On the other hand, if $a \in C_j$ for some j then the d-element $\{a\} \cup (A - B_j)$ pins all members of $\mathcal{G}^i$ in $D - \{a\}$. If, however, $a \in A$ then any d-element $E \subset A$ containing a pins the members of $\mathcal{G}^i$ in $D - \{a\}$. Therefore D is really a member of $\mathcal{G}^i(d)$. It is easy to see that there is no other member.

Choose $i = \left\lfloor k(1 - \frac{1}{d}) \right\rfloor$. Then the size of D, given by the left hand side of (3.1) asymptotically becomes

$$\frac{(d-1)^{d-1}}{d^d (d-1)!} k^d.$$

(3.1) gives a condition how large n has to be. □

Upper estimate. Let $G \in \mathcal{F}(d)$ where $\mathcal{F} \subset \binom{X}{\le k}$ (the latter one denotes the family of all subsets of X of size at most k). We will prove that $|G| \le dk^d$. Since we have to consider only the subsets of G, so it can be supposed that all members of $\mathcal{F}$ are subsets of G.

Proposition 3.1 defines d-element subsets D of G each of them is pinning $\mathcal{F}$. Moreover, still by Proposition 3.1, their union is G. Denote this family by $\mathcal{D}$. We know

$$\cup_{D \in \mathcal{D}} D = G, \tag{3.2}$$

$$D \cap F \ne \emptyset \quad \text{for all} \quad D \in \mathcal{D}, F \in \mathcal{F} \tag{3.3}$$

and $\mathcal{F}$ cannot be pinned by a set with less than d elements.

Let $I \subseteq G$. Define the I-degree of $\mathcal{D}$ as the number of members of $\mathcal{D}$ containing I, that is,

$$\deg_I(\mathcal{D}) = |\{D \in \mathcal{D} : \ I \subset D\}|.$$

Lemma 3.2 *If $|I| < d$ then*

$$\deg_I(\mathcal{D}) \leq k^{d-|I|}.$$

Proof. We use induction on $j = d - |I|$. Suppose that $j = d - |I| = 1$, that is, $|I| = d-1$. If all members of $\mathcal{F}$ meet I then $\mathcal{F}$ can be pinned by $d-1$ elements, a contradiction. Therefore there is an $F \in \mathcal{F}$ which is disjoint to I. By (3.3) all the sets D satisfying $I \subset D$ must intersect this F, therefore their number is $\leq |F| \leq k$. This case is settled.

Now suppose that the statement is true for $j = d-|I| \geq 1$ and prove it for $j+1 = d-|I|$. Let $|I^*| = d-j-1$. There must exist an $F \in \mathcal{F}, F \cap I^* = \emptyset$ otherwise $\mathcal{F}$ is pinned by less than d elements, a contradiction. Let $F = \{x_1, \ldots, x_l\}$ where $l \leq k$. By (3.3) we have

$$\{D \in \mathcal{D} : \ I^* \subset D\} = \cup_{i=1}^{l} \{D \in \mathcal{D} : \ (I^* \cup \{x_i\}) \subset D\}. \tag{3.4}$$

The sizes of the sets on the right hand side are $\deg_{I^* \cup \{x_i\}}(\mathcal{D})$ which are at most $k^{d-|I^*|-1} = k^j$ by the induction hypothesis. Using (3.4)

$$\deg_{I^*}(\mathcal{D}) \leq lk^{d-|I^*|-1} \leq k^{d-|I^*|}$$

is obtained, proving the lemma. □

Finally, consider any $F = \{y_1, \ldots, y_r\} \in \mathcal{F}$ where $r \leq k$. By (3.3), the families $\{D \in \mathcal{D} : y_i \in D\}$ cover $\mathcal{D}$. Apply the lemma for $I = \{y_i\}$:

$$|\{D \in \mathcal{D} : y_i \in D\}| \leq k^{d-1}.$$

This implies $|\mathcal{D}| \leq k^d$ and

$$|\cup_{D \in \mathcal{D}} D| \leq |\mathcal{D}|d \leq dk^d.$$

Application of (3.2) completes the proof: $|G| \leq dk^d$. □

Proof of Theorem 2.4 Let $d = 2e+1$. Apply the results of Theorem 2.3 first for a family $\mathcal{F}_a$. Since its members are not larger than k, the theorem

implies that all members of $\mathcal{F}_a(2e+1)$ are of size at most $c_1 k^{2e+1}$. Since this is true for every a, the union of the families has the same property.

On the other hand, take the inclusion-free family $\mathcal{F}$ giving the optimum in the lower estimation in Theorem 2.3. Lemma 2.2 defines a system of functional dependencies (database) in which $\mathcal{F}_a = \mathcal{F} \cup \{a\}$ holds for some $a \in \Omega$. Therefore $\mathcal{F}(2e+1)$ contains sets of size at least $c_2 k^{2e+1}$. □

4 Further problems

1. Although Theorem 2.3 determines the order of magnitude of the smallest size in the "worst" family, it does give the exact value. We believe that the lower estimate is sharp, our construction is the best possible.

Conjecture 4.1 *If* $\mathcal{F} \subseteq \binom{X}{\leq k}$ *then* $\mathcal{F}(d)$ *has a member with size at most*

$$\max_i \{i + d - 1 + \binom{i+d-1}{i}(k-i)\}.$$

for $n_0(k,d) \leq n$.

2. Can the systems of e-error-correcting dependencies be characterized?

3. The following problem sounds similar to the problem treated here, but it is actually very different. Suppose that the data go through a noisy channel, where each data can be distorted with a small probability. So, in this case only M^* is known, no information is available on the structure of M. What is the relationship between the "functional dependencies" found in M^* and the real functional dependencies in M? This is the real problem arising in data mining in a distorted database.

References

[1] Armstrong, W.W., Dependency structures of data base relationship, in: *Information Processing 74*, North-Holland, Amsterdam, pp. 580-583.

[2] Demetrovics, J., On the equivalence of candidate keys with Sperner systems, *Acta Cybernet.* **4**(1979) 247-252.

[3] Demetrovics J., Füredi, Z, and Katona, G.O.H.: Minimum matrix representation of closure operations, *Discrete Appl. Math.* **11**(1985) 115-128.

[4] Demetrovics J., G.O.H. Katona, Miklós, D.: Error-correcting keys in relational databases, in *Foundations of Informationn and Knowledge Systems, FoIKS 2000* (K.-D. Schewe and B. Thalheim eds.) Lecture Notes in Computer Science, **1762**, Springer, 2000, pp. 88-93.