Random databases with correlated data

Gyula O.H. Katona
Rényi Institute, Budapest, Hungary
ohkatona@renyi.hu

Dedicated to Professor
Bernhard Thalheim
for his 60th birthday

1 Introduction

Consider the data of a class in a school (in Europe), it can be supposed that the last name is a key, that is all other data are functionally dependent on it. Considering the whole school, the probability of having two students with the same last name is pretty high, so the last name cannot be taken as a key. But, very likely the first and last names together from a key. It will be certainly not true for the data of a large city.

The example above illustrates that, considering the database to be random, the size (number of rows) largely determines which functional dependencies can be considered valid. The aim of the present paper is to give a model of this situation. The first attempts in this directions were the papers of Demetrovics, Katona, Miklós, Seleznjev and Thalheim [1], [2]. There the authors supposed that the data of one individual are probabilistically independent. It was shown even in this case that a set of *constant times the logarithm of the size of the database* many columns will functionally determine a given other column with high probability. Their model however was not able to include "real" functional dependencies or situations like "very probably functionally dependent". The aim of the present paper is to extend the results in this direction.

Let Ω be the set of attributes, $|\Omega| = n$. The set of all possible entries is denoted by E. (If the distinct attributes have different sets of entries

then E is their union.) Let one row of the database is the random vector $(\xi_1, \xi_2, \ldots, \xi_n)$ where the ξs are not necessarily independent, the distribution is given by the probabilities

$$\Pr(\xi_1 = u_1, \xi_2 = u_2, \ldots, \xi_n = u_n) \tag{1}$$

for all possible entries $u_1, u_2, \ldots, u_n \in E$. Let $A \subset \Omega, b \in \Omega$. We say that b *functionally depends on A with probability one* if the probabilities

$$\Pr(\xi_i = u_i (i \in A), \xi_b = u_b)$$

are zero for all but one $u_b \in E$ for any choice of entries $u_i \in E (i \in A)$. On the other hand the individuals, the rows are chosen independently. In terms of probability theory, consider m (totally) independently chosen realizations of the random vector whose probabilities are determined by (1).

An entropy like function is needed to our further investigations. Let ξ and η be two, not necessarily independent random variables. The probability of the event that $\xi = k$ and $\eta = \ell$ is $p_{k,\ell}$, the probability of ξ being k is $p_k = \sum_\ell p_{k\ell}$. Define

$$H_2(\xi \to \eta) = -\log_2 \left(\sum_k p_k^2 - \sum_{k,\ell} p_{k,\ell}^2 \right). \tag{2}$$

This quantity is related to the Rényi entropy of order 2 (see [4] and [5]).

Let $A \subset \Omega, b \in \Omega, b \notin A$. The random vector of the coordinates $\xi_i (i \in A)$ will be denoted by α. The probability of the event that α is equal to the kth sequence is denoted by $p_k(A)$. Moreover, the probability of the event that α is equal to the kth sequence and ξ_b has the ℓth entry is $p_{k,\ell}(A, b)$. Our crucial notion is defined in the following way:

$$H_2(A \to b) = H_2(\alpha \to \xi_b). \tag{3}$$

The Heuristic Version of the Theorem. *The functional dependency $A \to b$ "seems to hold" (there are no two rows equal in the entries belonging to A and different in the column of b) with large probability in a random database of size m if and only if $2 \log_2 m$ is much smaller than $H_2(A \to b)$.*

The statement above will be made more clear by analyzing two special cases. First let us suppose that all the ξ's are independent in (1) and each

of them has a probability distribution $(q_1, q_2, \ldots, q_R)$. Then the probabilities in question are

$$\Pr(\xi_i = u_i (i \in A), \xi_b = u_b) = \prod_{i \in A} q_{u_i} \cdot q_{u_b}.$$

These probabilities will pay the role of $p_{k\ell}$ in (2), while p_k will be

$$\Pr(\xi_i = u_i (i \in A)) = \prod_{i \in A} q_{u_i}.$$

It is easy to see that in this case (deleting the arguments A and b)

$$\sum_k p_k^2 - \sum_{k,\ell} p_{k\ell}^2 = \sum_k p_k^2 - \sum_k p_k^2 \sum_\ell r_\ell^2 = \left(\sum_k p_k^2 \right) - \left(1 - \sum_\ell r_\ell^2 \right). \quad (4)$$

On the other hand,

$$\sum_\ell r_\ell^2 = \left(\sum_i q_i^2 \right)^{|A|}. \quad (5)$$

Using (2), (3), (4) and (5) we obtain

$$H_2(A \to b) = |A| \cdot H_2(q_1, q_2, \ldots q_R) - \log_2 \left(1 - \sum_\ell q_\ell^2 \right)$$

where $H_2(q_1, q_2, \ldots q_R) = \log_2 \sum_i q_i^2$ is the Rényi entropy of order 2 ([4], [5]). Here the first term tends to infinity with $|A|$ while the second term is constant. $H_2(A \to b)$ is close to $|A| \cdot H_2(q_1, q_2, \ldots q_R)$. The Theorem means in this case that $A \to b$ holds for a random database of size m if $2 \log_2 m$ is less than $|A| \cdot H_2(q_1, q_2, \ldots q_R)$, that is, for the As satisfying $|A| > \frac{2 \log_2 m}{H_2(q_1, q_2, \ldots q_R)}$. This was proved in [2].

The other important special case is when b is really functionally dependent on A. Then $p_{k\ell} = p_k$ for a uniquely determined $\ell = \ell(k)$, all other $p_{k\ell}$s are zero. Therefore the last term in (2) is equal to $\sum_k p_k^2$, (2) is plus infinity. The Theorem says in this case that $A \to b$ holds when $2 \log_2 m$ is less than ∞, that is always.

2 The exact forms of the Theorem

It will be supposed that the database consists of m (totally) independently chosen rows of the random vector defined by the probability distribution (1). Our result is of asymptotic nature, it is valid for large matrices, large number of columns and rows. More precisely we will assume that $n = |\Omega|, |A|$ depend on m what tends to infinity. It may seem more natural to take n to be the main variable and to suppose that the other quantities depend on it while it tends to infinity. However the size of the asymptotical existence of $A \to b$ is independent on n it only depends on the relation of m and $H_2(A \to b)$. This is why it is better to consider m as the basic variable.

It will be supposed that the distribution (1) for n' is the "continuation" of the one for n, that is, the probabilities in (1) can be obtained by summing the probabilities for n' for $\xi_{n+1}, \ldots, \xi_{n'}$. The column b is fixed, while $|A|$ tends to infinity by adding newer and newer columns (distinct from b) to A.

Some more definitions are needed to the formulation of the theorem. The probability of the of the event that $A \to b$ $(A \subset \Omega, b \in \Omega)$ holds in a database of size m is denoted by $\Pr(A \to b, m)$. Let $p(\alpha, \xi_b, I)$ denote the probability of the event that the pair of two independent copies $(\alpha_1, \xi_{b,1}), (\alpha_2, \xi_{b,2})$ gives a counter-example, that is, $\Pr(\alpha_1 = \alpha_2, \xi_{b,1} \neq \xi_{b,2})$. Similarly $p(\alpha, \xi_b, V)$ denotes the probability of the event that the triple $(\alpha_1, \xi_{b,1}), (\alpha_2, \xi_{b,2}), (\alpha_3, \xi_{b,3})$ gives two counter-examples in the following way: $\alpha_1 = \alpha_2 = \alpha_3, \xi_{b,1} \neq \xi_{b,2} \neq \xi_{b,3}$. Finally $p(\alpha, \xi_b, N)$ is the probability of the event that the quadruple $(\alpha_1, \xi_{b,1}), (\alpha_2, \xi_{b,2}), (\alpha_3, \xi_{b,3}), (\alpha_4, \xi_{b,4})$ gives three counter-examples forming a path: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4, \xi_{b,1} \neq \xi_{b,2} \neq \xi_{b,3} \neq \xi_{b,4}$.

The first exact form of the Theorem is a repetition/implementation of the main theorem in [3]. This theorem is stated for two random variables. The only novelty here is that one of these variables is a random vector α. But this causes no real change. Therefore the theorem below needs no proof here. The interested reader is referred to [3].

Theorem 1.

$$\Pr(A \to b, m) \to \begin{cases} 0 & \text{if } 2\log_2 m - H_2(A \to b) \to +\infty, \\ e^{-2^{a-1}} & \text{if } 2\log_2 m - H_2(A \to b) \to a, \\ 1 & \text{if } 2\log_2 m - H_2(A \to b) \to -\infty. \end{cases}$$

under the assumptions that

$$\frac{p(\alpha, \xi_b, V)^2}{p(\alpha, \xi_b, I)^3} \to 0 \tag{6}$$

and

$$\frac{p(\alpha, \xi_b, N)}{p(\alpha, \xi_b, I)^2} \to 0 \tag{7}$$

hold.

Although this is the most general form of the statement, known to us, it is difficult to check if the conditions (6) and (7) hold. However, exploiting the matrix structure in this case we can give weaker, but more natural conditions. Let $p_\kappa(A)$ denote the probability of the event that $\alpha = \kappa$. Moreover, $p_{\kappa,\ell}(A, b)$ denotes the probability of the event that $\alpha = \kappa, \xi_b = \ell$.

Theorem 2.

$$\Pr(A \to b, m) \to \begin{cases} 0 & \text{if } 2\log_2 m - H_2(A \to b) \to +\infty, \\ e^{-2^{a-1}} & \text{if } 2\log_2 m - H_2(A \to b) \to a, \\ 1 & \text{if } 2\log_2 m - H_2(A \to b) \to -\infty. \end{cases}$$

under the following assumptions:

(i)

$$\frac{\max_\kappa p_\kappa(A)}{\sqrt{\sum_\kappa p_\kappa^2(A)}} \to 0, \tag{8}$$

(ii) There is a constant $0 < u < 1$ *independent of* A *such that*

$$\frac{\sum_{\kappa,\ell} p_{\kappa,\ell}^2(A, b)}{\sum_\kappa p_\kappa^2(A)} \leq u \tag{9}$$

hold.

Proof. We have to prove that (i) and (ii) imply both (6) and (7). Let us start with some elementary lemmas. The first two of them prove that if a sequence of non-negative numbers is given with a fixed sum and an upper bound c is given on them then the sum of their squares is maximized for a choice with (one exception) all members $= c$ or 0. We give the proof for sake of completeness.

Lemma 1. *Let the real numbers* $0 \leq a, b, c$ *satisfy the inequalities* $b \leq a \leq c \leq a + b$. *Then*

$$a^2 + b^2 \leq c^2 + (a + b - c)^2 \tag{10}$$

holds.

Proof. Consider the function $x^2 + (a+b-x)^2$. It is increasing from $\frac{a+b}{2}$. The conditions of the lemma imply (10), considering the values $x = a$ and $x = c$. □

Lemma 2. *Let $a_1, a_2, \ldots, a_N$ be non-negative real numbers with sum $\sum_i a = s$. If $a_i \leq c$ holds for all $1 \leq i \leq N$ then*

$$\sum_i a_i^2 \leq \left\lceil \frac{s}{c} \right\rceil c^2 \tag{11}$$

is true.

Proof. We use induction over N. The case $N = 1$ is trivial. Order the numbers in the following way: $a_1 \leq a_2 \leq \ldots \leq a_N \leq c$. If $a_N = c$ then delete this member and use the inductional hypothesis. Otherwise, if $a_N < c$ two cases will be distinguished. Firstly, if $c \leq a_{N-1} + a_N$ then apply Lemma 1 with $a = a_N, b = a_{N-1}$. Replacing a_N by c and a_{N_1} by $a_{N-1} + a_N - c$ a new sequence of numbers is obtained with the the same sum and non-decreased sum of squares. It is sufficient to prove the statement for this sequence, but this follows from the previous case, since it contains $a_N = c$. Secondly, if $c > a_{N-1} + a_N$ then apply Lemma 1 with $b = a_{N-1}, a = a_N, c = a_{N-1} + a_N$. The so obtained inequality, $a_{N-1}^2 + a_N^2 \leq (a_{N-1} + a_N)^2 + 0^2$ (what can be directly seen) shows that replacing a_{N-1} and a_N by $a_{N-1} + a_N$ and 0 the sum is unchanged, the sum of the squares is non-decreased. Since this sequence contains a 0, omitting this the induction can be used, again. □

Lemma 3. *Let $q_1, q_2, \ldots, q_N$ be non-negative real numbers, where all of these (including N) depend on n what tends to the infinity. Then*

$$\frac{\max_k q_k}{\sum_k q_k} \to 0 \tag{12}$$

implies

$$\frac{\sum_k q_k^2}{\left(\sum_k q_k\right)^2} \to 0. \tag{13}$$

Proof. Use (11) of Lemma 2 with $c(n) = \max_k q_k$ and $s = \sum_k q_k$:

$$\frac{\sum_k q_k^2}{\left(\sum_k q_k\right)^2} \leq \frac{\left(\frac{\sum_k q_k}{c(n)} + 1\right) c^2(n)}{\left(\sum_k q_k\right)^2} = \frac{c(n)}{\sum_k q_k} + \left(\frac{c(n)}{\sum_k q_k}\right)^2$$

shows that (12) really implies (13). □

Return to the proof of Theorem 2. Lemma 3 will be applied for the values $p_\kappa^2(A)$ in place of q_i. Condition (12) becomes exactly (8), therefore (13) gives

$$\frac{\sum_\kappa p_\kappa^4(A)}{\left(\sum_\kappa p_\kappa^2(A)\right)^2} \to 0. \tag{14}$$

Let us now give a lower estimate on $p(\alpha, \xi_b, I)$ using (ii) (that is (9))

$$p(\alpha, \xi_b, I) = \sum_\kappa p_\kappa^2(A) - \sum_{\kappa,\ell} p_{\kappa,\ell}^2(A,b) \geq (1-u) \sum_\kappa p_\kappa^2(A). \tag{15}$$

Recall that $p(\alpha, \xi_b, N)$ is the probability of the event that the quadruple $(\alpha_1, \xi_{b,1}), (\alpha_2, \xi_{b,2}), (\alpha_3, \xi_{b,3}), (\alpha_4, \xi_{b,4})$ gives three counter-examples forming a path: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4, \xi_{b,1} \neq \xi_{b,2} \neq \xi_{b,3} \neq \xi_{b,4}$. This is a subevent of the event that $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$. The probability of the latter one is $\sum_\kappa p_\kappa^4(A)$. Hence we have

$$p(\alpha, \xi_b, N) \leq \sum_\kappa p_\kappa^4(A). \tag{16}$$

(15) and (16) give an upper bound on the left hand side of (7):

$$\frac{p(\alpha, \xi_b, N)}{p(\alpha, \xi_b, I)^2} \leq \frac{\sum_\kappa p_\kappa^4(A)}{(1-u)^2 \left(\sum_\kappa p_\kappa^2(A)\right)^2}. \tag{17}$$

The right hand side tends to 0 by (14), proving (7).

The left hand side of (6) can be similarly upperbounded:

$$\frac{p(\alpha, \xi_b, V)^2}{p(\alpha, \xi_b, I)^3} \leq \frac{\left(\sum_\kappa p_\kappa^3(A)\right)^2}{(1-u)^3 \left(\sum_\kappa p_\kappa^2(A)\right)^3}. \tag{18}$$

Apply the well-known Cauchy-Bunyakovsky-Schwarz inequality

$$\left(\sum_i a_i b_i\right)^2 \leq \left(\sum_i a_i^2\right) \left(\sum_i b_i^2\right)$$

with $p_\kappa(A)$ and $p_\kappa^2(A)$:

$$\left(\sum_\kappa p_\kappa^3(A)\right)^2 \leq \left(\sum_\kappa p_\kappa^2(A)\right) \left(\sum_\kappa p_\kappa^4(A)\right).$$

This latter inequality implies

$$\frac{\left(\sum_\kappa p_\kappa^3(A)\right)^2}{\left(\sum_\kappa p_\kappa^2(A)\right)^3} \le \frac{\sum_\kappa p_\kappa^4(A)}{\left(\sum_\kappa p_\kappa^2(A)\right)^2}. \tag{19}$$

(18), (19) and (14) prove (6). □

3 Remarks, future work

Related earlier work. In addition to the papers [1], [2], mentioned in the introduction, one should mention the important works of Seleznjev and Thalheim [6], [7] on the probabilistic-statistical properties of databases.

On the conditions of Theorem 2. Condition (i) is a rather weak one, it is satisfied in a very wide range. However it is stronger than the condition $\max_\kappa p_\kappa(A) \to 0$. An example when the latter one holds but (i) does not is the following. Let N denote the total number of members (probabilities), and choose the largest one to be $\frac{1}{\sqrt{N}}$, the other ones are equal, and add up to 1. Then the limit in (8) is $\frac{1}{2}$, not 0.

On the other hand, condition (ii) (that is (9)) is strong. It excludes *e.g.* the case when b is "very probably functionally dependent" on A. We are sure that (6) and (7) can be proved under (8) and a much weaker condition than (9). It needs a deeper analysis of the situation. For instance, the rough estimate (16) is not sufficient in this more general case.

Future work. Besides the analytic work suggested in the previous paragraph, one should consider a more general setting of the whole problem. Already the present setting has a certain "data mining" nature. We investigated here that when (at what size?) a hidden, weak statistical dependence becomes visible. A possible more general setting is when the following question is investigated. Given the statistical dependence a certain number of examples can be expected. At what size have we at least (say) half of this number. Another possible step forward if the "quality" of the examples is also taken into consideration.

References

[1] Demetrovics, J., Katona, G.O.H., Miklós, D., Seleznjev, O., Thalheim, B., Asymptotic properties of keys and functional dependencies in random databases, *Theoretical Computer Sciences* **190**(1998) 151-166.

[2] Demetrovics, J., Katona, G.O.H., Miklós, D., Seleznjev, O., Thalheim, B., Functional dependencies in random databases, *Studia Sci. Math. Hungar.* **34**(1998) 127-140.

[3] Gyula O.H. Katona, Testing functional connection between two random variables, Prokhorov Festschrift, accepted.

[4] Rényi Alfréd, Some fundamental questions of information theory (in Hungarian), *MTA III Oszt. Közl.* **10**(1960) 251-282.

[5] Rényi Alfréd, On measures of information and entropy, *Proc. of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, 1960*, 1961, pp. 547-561.

[6] Seleznjev, Oleg, Thalheim, Bernhard, Average Case Analysis in Database Problems, *Methodology and Computing in Applied Probability* **5**(4)(2003) 395-418.

[7] Seleznjev, Oleg, Thalheim, Bernhard, Random Databases with Approximate Record Matching, *Methodology and Computing in Applied Probability* **12**(1)(2010) 63-89.