New results in cutting seriation for approximate #SAT

Israël César Lerman[1] and Valérie Rouat[2]

[1] IRISA, Campus de Beaulieu, F-35042 Rennes Cedex, France – lerman@irisa.fr
[2] CELAR, F-35170 Bruz, France – rouat@celar.fr

Summary. The general problem discussed here concerns the approximation of the number of solutions of a boolean formula in conjunctive normal form F. Results previously obtained (Rouat (1999), Lerman and Rouat (1999)) are reconsidered and completed. Our method is based on the general principle "divide to resolve". The division is achieved by cutting a seriation built on an incidence data table associated with F. In this, the independence probability concept is finely exploited. Theoretical justification and intensive experimentation validate the considerable reduction of the computational complexity obtained by our method.

1 Introduction

The Classification and Combinatorial Data Analysis methods have two general and related aims:

1. extracting high density regions in the data representation space;
2. reducing the complexity of the data interpretation.

This objective can be associated with fundamental problems in the field of computational complexity (Lerman (1995), Rouat (1999), Lerman and Rouat (1999)). The most representative of them concern satisfiability of boolean equations. Let us introduce these problems; namely the SAT problem and the #SAT problem.

Consider a set $V = \{x_1, \ldots, x_i, \ldots, x_N\}$ of boolean variables, a clause built on V is a disjunction of literals $y_1 \vee y_2 \vee \ldots \vee y_q \vee \ldots \vee y_r (r < N)$ where $\{y_1, y_2, \ldots, y_q, \ldots, y_r\}$ is defined from a subset of r variables of V. Each of them is taken in its positive or (exclusively) negative form (x_i or $\neg x_i$). An assignment of the boolean variables satisfies the clause if at least one of the variables y_q is true, $1 \leq q \leq r$.

A SAT instance is defined by a conjunction of clauses built on the set V of boolean variables. The SAT problem is that of the satisfiability of a SAT instance; that is to say, the recognition of the existence of a solution. In other words, does there exist an assignment of the boolean variables for which the SAT instance is true?

#SAT problem consists of evaluating the number of solutions of a SAT instance. Obviously an answer to the SAT problem is immediately provided by a solution of the #SAT problem.

SAT problem is at the origin of the definition of the NP-complete problems (Cook (1971)). They constitute the most difficult subclass of the "Non deterministic Polynomial problems" (NP-problems). The nature of a large class of decision problems is NP. Establishing the conjecture for which there does not exist a Polynomial algorithm to resolve the SAT problem (P $\neq$ NP) is the most crucial point in the computational complexity theory. The #SAT problem for which all the solutions of a SAT instance have to be enumerated is clearly and a priori more difficult. Indeed it is situated in the class of #P-complete problems (Valiant (1979)). This class "comprises" all the polynomial hierarchy (Toda (1989)). It has then very particular importance in the computational complexity theory. Consequently, the part of #SAT problem in this theory is becoming more and more important these last year (Papadimitriou (1994)).

Several algorithms of exact resolution of #SAT have been proposed (Dubois (1991), Lozinskii (1992)). But all of them have an exponential nature and then become intractable even for reduced sizes of the SAT instances. However, approximating the number of solutions has a crucial interest in the field of computational complexity. Very important applications arise in computing probabilities managing large systems. Many methods have been proposed. One approach consists in interrupting an exact resolution algorithm in order to infer an estimation (Rouat and Lerman (1997, 1998)). Other approaches are based on random sampling in a representation space of the SAT instance (Karp and Luby (1983), Bailleux and Chabrier (1997)).

The basic idea of our method is associated with the general principle "divide to resolve". The matter consists of dividing the whole problem into two subproblems of similar size and reconstituting — in a polynomial algorithm — an approximate evaluation of the global solution, from the exact solutions of the two subproblems. For this purpose an incidence data table crossing clauses by variables is associated with the SAT instance. In these conditions our method can be divided into three phases:

1. application of a specific technique of seriation;
2. cutting the seriation into two connected parts of comparable sizes, in an optimal manner with respect to a statistical independence criterion, having a polynomial cost;
3. reconstitution of an approximate value of the total number of solutions of the whole instance by means of a relevant formula, taking into account the nature of the random simulation of the SAT instance.

Precisely, we reconsider in this paper, by means of a new reconstitution formula, the experimental results obtained previously (Rouat (1999), Lerman and Rouat (1999)). The new equation takes more intimately into account the statistical dependency of the above mentioned two segments of the seriation.

All of our experiments concern the classical cases of random 3SAT and random 2SAT. The generation model assumes total probabilistic independence and uniform distribution over the clause space, for which exactly three (resp. two) variables per clause are instanciated for 3SAT (resp. 2SAT). Thus, for this model the probability of a given clause is $1/\binom{N}{3}2^3$ (resp. $1/\binom{N}{2}2^2$), where $\binom{N}{k}$ denotes a binomial coefficient. This random model for which there is not any hidden statistical structure, provides the difficult cases of algorithmic resolution. It has been shown that the hardest of them for 3SAT corresponds to a ratio of 1.2 between the number P of clauses and the number N of variables (Rouat (1999)).

2 The representation of the problem

2.1 Preliminary definitions

Let us recall the combinatorial and geometrical representation (Lerman (1995), Rouat and Lerman (1997, 1998)) that we have introduced and exploited in our analysis. This representation allows Combinatorial Data Analysis to have a part in treating SAT problems.

As mentioned above relative to a set $V = \{x_1, \ldots, x_i, \ldots, x_N\}$ of boolean variables, a clause of order $r(r < N)$ can be written:

$$C^r = y_1 \vee y_2 \vee \ldots \vee y_q \vee \ldots \vee y_r \tag{1}$$

where $\{1, 2, \ldots, q, \ldots, r\}$ designates a subset $\{i_1, i_2, \ldots, i_q, \ldots, i_r\}$ of r subscripts of $\{1, 2, \ldots, i, \ldots, N\}$ and where y_q represents x_{i_q} or (exclusively) $\neg x_{i_q}$, $1 \leq q \leq r$. Thus, for example, by supposing N greater than 7,
$C^3 = x_1 \vee \neg x_3 \vee x_7$ is a clause of order 3 for which $y_1 = x_1, y_2 = \neg x_3$ and $y_3 = x_7$.

An assignment of the boolean variables is a solution of (or satisfies) the clause if and only if at least one of the variables $y_q, 1 \leq q \leq r$, is true. It is the case in the preceding example if x_1 is true or (non exclusively) x_3 is false or (non exclusively) x_7 is true.

2.2 Pinpoint cylinder associated with a clause

A logical cube $\{0,1\}^N$ is associated with the set V of the N boolean variables. It corresponds to the value set of the vector of boolean variables $(x_1, \ldots, x_i, \ldots, x_N)$. A value 1 (resp. 0) of the ith component does mean that the variable x_i is true (resp. false), $1 \leq i \leq N$. The pinpoint cylinder associated with a given clause C is simply defined by the set of points of $\{0,1\}^N$ which falsify the clause C. This set of vertices has a particular geometric structure. More precisely, associate with the above clause C^r, its negation, the anti-clause $\neg C^r$. By writing C^r in a more explicit form

$$C^r = y_{i_1} \vee y_{i_2} \vee \ldots \vee y_{i_q} \vee \ldots \vee y_{i_r}, \tag{2}$$

we have $$\neg C^r = \neg y_{i_1} \wedge \neg y_{i_2} \wedge \ldots \wedge \neg y_{i_q} \wedge \ldots \wedge \neg y_{i_r} \tag{3}$$
where $\neg y_{i_q} = x_{i_q}$ (resp. $\neg x_{i_q}$) if $y_{i_q} = \neg x_{i_q}$ (resp. x_{i_q}), $1 \leq q \leq r$.

The subset of points of the cube $\{0,1\}^N$ satisfying (3) may be represented by a vector of which the only specified components are $i_1, i_2, \ldots, i_q, \ldots, i_r$, the other components being indeterminate. More precisely, by denoting α_i the ith component of such a vector, we have:

$$\alpha_{i_q} = 1 \text{ (resp. 0) if } y_{i_q} = \neg x_{i_q} \text{ (resp. } x_{i_q}\text{)}, 1 \leq q \leq r,$$
$$\alpha_i = \varepsilon \text{ if } i \notin \{i_1, i_2, \ldots, i_q, \ldots, i_r\}$$

where ε is an indeterminate boolean.

This structure that we can denote by

$$E^r = E(C^r) = \{(i_1, i_2, \ldots, i_r), (\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_r})\} \tag{4}$$

defines in the logical cube space a cylinder whose the basis is the point $(\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_r})$ in the subspace underlined by the components $i_1, i_2, \ldots, i_r$. This is the reason why we call E^r a "pinpoint cylinder of order r". Note that intersection of pinpoint cylinders is a pinpoint cylinder.

2.3 Set theoretic expressions for SAT and #SAT problems

SAT instance can be put in the following conjunctive normal form

$$F = C_1^{r_1} \wedge C_2^{r_2} \wedge \ldots \wedge C_i^{r_i} \wedge \ldots \wedge C_p^{r_p} \tag{5}$$

where $C_i^{r_i}$ is a clause of order $r_i, 1 \leq i \leq P$. Let $E_i^{r_i}$ denote the pinpoint cylinder associated with $C_i^{r_i}, 1 \leq i \leq P$. The negated F formula $\neg F$ will be represented by the union:

$$G = \bigcup_{1 \leq i \leq P} E_i^{r_i}. \tag{6}$$

In these conditions, the SAT instance is satisfiable if and only if G is a strict subset of the cube $\{0,1\}^N$; that is to say, if and only if G does not cover all the cube ($2^N - \text{card}(G) \neq 0$). On the other hand, the #SAT problem consists of evaluating the cardinality $\text{card}(G)$ of G. Thus, the resolution of the #SAT problem will be given by:

$$\text{NBS}(F) = 2^N - \text{card}(G). \tag{7}$$

3 Logical and probabilistic independences between two SAT instances on the same variable set

According to (Simon and Dubois (1989)) two clauses C and C' are "logically independent" if and only if no assignment of the N variables contradicts both clauses. Since, the contradiction of a given clause is equivalent to the satisfiability of the associated anti-clause, this notion of logical independence corresponds exactly to disjunction in the set theoretic sense, between the two pinpoint cylinders $E(C)$ and $E(C')$ respectively associated with C and C':

$$\text{logical independence between } C \text{ and } C' \Longleftrightarrow E(C) \cap E(C') = \emptyset. \tag{8}$$

Let us now denote by C and C' the respective sets of clauses corresponding to two SAT instances. More explicitly we have:

$$\mathsf{C} = \{C_i^{r_i} | 1 \leq i \leq r\} \quad \text{and} \quad \mathsf{C'} = \{C'^{r'_i}_i | 1 \leq i \leq r\}. \tag{9}$$

Consider now the sets $G(\mathsf{C})$ and $G(\mathsf{C'})$ respectively associated with C and $\mathsf{C'}$ in the same manner as G has been associated with F (see (5) and (6) above). $G(\mathsf{C})$ and $G(\mathsf{C'})$ are unions of pinpoint cylinders. It follows that we may extend the above definition (see (8)) by putting:

$$\text{logical independence between } \mathsf{C} \text{ and } \mathsf{C'} \Longleftrightarrow G(\mathsf{C}) \cap G(\mathsf{C'}) = \emptyset. \tag{10}$$

The second member is equivalent to:

$$\forall (i,j), 1 \leq i,j \leq P, \quad C_i^{r_i} \cap C'^{r'_j}_j = \emptyset. \tag{11}$$

However the concerned independence notion that we have studied and exploited (Rouat (1999), Lerman and Rouat (1999)) is probabilistic. New results are reported here. Let us now recall the general theoretical framework.

For a pinpoint cylinder $E(C)$ representing a clause C, we define the probability $\mathrm{P}[E(C)]$ for a vertex — taken randomly in the cube $\{0,1\}^N$, provided by an uniform distribution — to enter $E(C)$. This probability represents the proportion of the cube vertices which belong to $E(C)$. Therefore, clearly $\mathrm{P}[E(C^r)] = 2^{-r}$, where C^r is a clause of order $r(r \leq N)$.

Definition 1. *The clauses C and C' are said to be independent (in probability) if and only if*

$$P[E(C) \cap E(C')] = P[E(C)] \times P[E(C')] \tag{12}$$

where $E(C)$ and $E(C')$ are the pinpoint cylinders associated with C and C'.

Let now $W(C)$ and $W(C')$ be two variable sets respectively instanciated in the clauses C and C', we have the following results (Rouat (1999), Lerman and Rouat (1999)).

Lemma 1. *The clauses C and C' are independent in probability if and only if $W(C)$ and $W(C')$ are disjoint ($W(C) \cap W(C') = \emptyset$).*

The generalization of the independence relation (12) to two sets of clauses C and $\mathsf{C'}$ (see (9) above) can be stated as follows:

$$\mathrm{P}[G(\mathsf{C}) \cap G(\mathsf{C'})] = \mathrm{P}[G(\mathsf{C})] \times \mathrm{P}[G(\mathsf{C'})] \tag{13}$$

where $G(\mathsf{C})$ and $G(\mathsf{C'})$ have been defined above. In these conditions we have the following:

Theorem 1. *C and $\mathsf{C'}$ are two independent sets of clauses if whatever the pair of clauses C and C' belonging respectively to C and $\mathsf{C'}$ ($C \in \mathsf{C}$ and $C' \in \mathsf{C'}$), C and C' are independent.*

Note that this condition is sufficient but not necessary. We also have the following result:

Theorem 2. *C and $\mathsf{C'}$ being two sets of clauses a necessary and sufficient condition for each clause of C to be independent of each clause of $\mathsf{C'}$ is that the variable sets $W(\mathsf{C})$ and $W(\mathsf{C'})$ are disjoint ($W(\mathsf{C}) \cap W(\mathsf{C'}) = \emptyset$).*

4 Measuring dependency degree and reconstitution formula

In the context of the principle of our method “divide to resolve” consider an ordered sequence of clauses $\mathsf{C} = \{C_i | 1 \leq i \leq P\}$ and imagine a decomposition of C into two sets A_c and B_c delimited by a given subscript c. More precisely,

$$\mathsf{A}_c = \{C_i | 1 \leq i \leq c\} \quad \text{and} \quad \mathsf{B}_c = \{C_i | c+1 \leq i \leq P\}. \tag{14}$$

Let

$$A_c = \{E_i | 1 \leq i \leq c\} \quad \text{and} \quad B_c = \{E_i | c+1 \leq i \leq P\} \tag{15}$$

the two sets of pinpoint cylinders respectively associated with A_c and B_c. Finally, denote by G_c and H_c the unions

$$G_c = \bigcup_{1 \leq i \leq c} E_i \quad \text{and} \quad H_c = \bigcup_{c+1 \leq i \leq P} E_i. \tag{16}$$

As mentioned above (see (13)) the probabilistic independence between A_c and B_c is expressed at the level of G_c and H_c by means of the equation:

$$\mathrm{P}[G_c \cap H_c] = \mathrm{P}[G_c] \times \mathrm{P}[H_c]. \tag{17}$$

In case of non independence the proposed dependency measure is given by the joint probability density with respect to the product of the marginal probabilities, namely:

$$\mathrm{dep}(G_c, H_c) = \frac{\mathrm{P}[G_c \cap H_c]}{\mathrm{P}[G_c] \times \mathrm{P}[H_c]}. \tag{18}$$

In these conditions the dependency degree between A_c and B_c is related to the distance of this index to unity.

Note that $G_c \cap H_c$ represents in its development the union of $c \times (P - c)$ pinpoint cylinders. Consequently, the computational complexity of $\mathrm{dep}(G_c, H_c)$ becomes exponential. To see that, refer to the inclusion-exclusion formula.

In order to avoid this complexity we have proposed an approximation of $\mathrm{dep}(G_c, H_c)$ given by:

$$\mathrm{apdep}(G_c, H_c) = \frac{\sum_{a \in G_c} \sum_{b \in H_c} \mathrm{P}[a \cap b]}{[\sum_{a \in G_c} \mathrm{P}[a]] \times [\sum_{b \in H_c} \mathrm{P}[b]]}. \tag{19}$$

It is of importance to note that this index preserves the basic properties of $\mathrm{dep}(G_c, H_c)$. As a matter of fact we have:

Theorem 3. *apdep(G_c, H_c) is equal to unity if G_c and H_c are independent in probability.*

Theorem 4. *apdep(G_c, H_c) is equal to zero if and only if G_c and H_c are disjoint.*

We leave the proofs of these theorems to be re-established by the reader.

It is clear that the computational complexity of $\mathrm{apdep}(G_c, H_c)$ is polynomial of order 2 with respect to the set C of clauses. Notice that all theses properties remain valid for $\mathrm{apdep}(G_c, H_c)^{\alpha}$ where α is a positive real number.

Let us now denote by I and J the respective formulas corresponding to G_c and H_c. On the other hand, note that the equation (7) can be written:

$$\mathrm{NBS}(F) = (1 - \mathrm{P}[F]) \times 2^N \tag{20}$$

where $\mathrm{P}[F]$ is the probability of non satisfiability of the formula F (see (5)). F is regarded here as the conjunction of I and J. Now, by considering the following formula

$$\mathrm{card}(G_c \cup H_c) = \mathrm{card}(G_c) + \mathrm{card}(H_c) - \mathrm{card}(G_c \cap H_c) \tag{21}$$

one may establish the following result:

Theorem 5.

$$\begin{aligned} NBS(F) = {} & \frac{NBS(I) \times NBS(J)}{2^N} \\ & + 2^N \left(1 - \frac{NBS(I)}{2^N}\right)\left(1 - \frac{NBS(J)}{2^N}\right)\left(dep(G_c, H_c) - 1\right). \end{aligned} \tag{22}$$

This equation can also be put in the following form:

$$\begin{aligned} \mathrm{NBS}(F) = {} & 2^N \times \frac{\mathrm{NBS}(I)}{2^N} \times \frac{\mathrm{NBS}(J)}{2^N} \times \\ & \left(1 + \left(1 - \frac{2^N}{\mathrm{NBS}(I)}\right)\left(1 - \frac{2^N}{\mathrm{NBS}(J)}\right)\left(\mathrm{dep}(G_c, H_c) - 1\right)\right). \end{aligned} \tag{23}$$

Imagine that G_c and H_c have the same size, namely $P/2$. In case of the considered random model the mathematical expectation of $\mathrm{NBS}(I)$ (resp. $\mathrm{NBS}(J)$) is given by[3] (Simon and Dubois (1989)) $(1 - 2^{-3})^{P/2} \times 2^N$ and then the magnitude order of the multiplicative factor of $(\mathrm{dep}(G_c, H_c) - 1)$ is given by $\left(1 - (8/7)^{P/2}\right)^2 \simeq (8/7)^P$ which is equal approximatively to $630,000$ for $P = 100$.

As described in the introduction our method consists in cutting the formula F into two complementary parts I and J having more or less the same number of clauses and as independent in probability as possible. For this decomposition for which with the above notations, we have $c \times (P - c)$ large and $\mathrm{dep}(G_c, H_c)$ near unity, a reconstitution formula is proposed. In our earlier experiments we have retained an approximation for which the complementary term added to 1 between the brackets of (23) is neglected. More precisely, the previous approximation that we denote by $\mathrm{ap1NBS}(F)$ is written:

$$\mathrm{ap1NBS}(F) = \frac{\mathrm{NBS}(I) \times \mathrm{NBS}(J)}{2^N}. \tag{24}$$

Now, for the new approximation, a part is given for the above mentioned complementary term by substituting $\mathrm{dep}(G_c, H_c)$ with $\mathrm{apdep}(G_c, H_c)^{\alpha}$, with a small value of α. The reason for this exponentiation consists of improving the approximation

[3] where I and J are conjunctions of clauses of order 3.

quality of $\mathrm{dep}(G_c,H_c)$ by $\mathrm{apdep}(G_c,H_c)$ and also numerical accuracy in computing. In these conditions, by denoting $\mathrm{ap2NBS}(F)$ the new approximation, we have:

$$\mathrm{ap2NBS}(F) = \frac{\mathrm{NBS}(I)\times\mathrm{NBS}(J)}{2^N} \\ +2^N\Big(1-\frac{\mathrm{NBS}(I)}{2^N}\Big)\Big(1-\frac{\mathrm{NBS}(J)}{2^N}\Big)\Big(\big(\mathrm{apdep}(G_c,H_c)\big)^{\alpha}-1\Big). \qquad (25)$$

5 “Divide to resolve”: cutting seriation associated with SAT

Let there be an arbitrary SAT instance for which we designate as previously by $\{x_1,x_2,\ldots,x_j,\ldots,x_N\}$ the variable set and by $\{C_1,\ldots,C_i,\ldots,C_P\}$ the clause set. The (i,j) entry of the incidence data table associated with such a SAT instance is defined by:

$$a_{ij} = \begin{cases} 0 \text{ if neither } x_j \text{ and } \neg x_j \text{ appear in the } i\text{th clause } C_i \\ 1 \text{ if } x_j \text{ or } \neg x_j \text{ appearsin the } i\text{th clause } C_i \end{cases} \qquad (26)$$

$1 \le i \le P$, $1 \le j \le N$.

According to the above theorem 2, the ideal structure to set up is of block seriation (Lerman (1972), Leredde (1979), Marcotorchino (1987)) with exactly two blocks having equal sizes. Clearly, this pure form is inaccessible in real cases of SAT instances. Moreover, it is quasi impossible for this form to occur in case of random SAT instances. Nevertheless and whatever will be the quality of the obtained result, we have to approximate as close as possible this form by permuting rows and columns of the incidence data table. We have shown (Rouat and Lerman (1997, 1998)) that the most flexible strategy consists of first determining global seriation on the incidence data table and second, cutting this seriation as optimally as possible. For this purpose, relative to the established seriation (see Figure 1) we define in the row set a median segment $[0.4P, 0.6P]$ covering 20% of the whole row set. In this segment we seek for the best cutting by means of the criterion $\mathrm{apdep}(G_c,H_c)$ (see (19)). More explicitly, in this latter expression, c indicates the row just before the cutting, G_c (resp. H_c) represents the union of the pinpoint cylinders respectively associated with the c first (resp. $(P-c)$ last) rows. Thus c is determined according to:

$$\mathrm{argmin}\Big\{\Big|\log\big(\mathrm{apdep}(G_c,H_c)\big)\Big| \quad \mid 0.4P \le c \le 0.6P\Big\}. \qquad (27)$$

A direct and specific technique of seriation has been proposed in the context of this research (Rouat and Lerman (1997, 1998), Rouat (1999)).

In these conditions, for each random generation of a SAT instance, the experimental design is decomposed as follows:

1. Seriation of the incidence data table defined in (26).
2. Determination of the best cutting according to the above criterion (27).
3. Computing the exact numbers of solutions of both sub-instances obtained by means of a variant of the Davis & Putnam algorithm (see in Davis and Putnam (1960), André and Dubois (1992)).

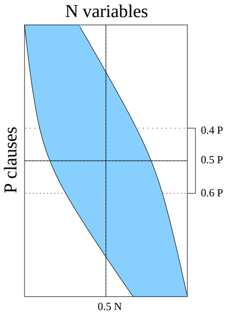

Fig. 1. Existence matrix: the clear part contains only zeros.

	2SAT		3SAT	
N α	70 10^{-5}	80 10^{-6}	40 10^{-4}	50 10^{-5}
ap1b ap2b	0.895 0.915	0.875 0.902	0.803 0.811	0.806 0.812
ap1 ap2	0.727 0.730	0.711 0.712	0.597 0.598	0.587 0.590

Table 1. Slopes of the regression lines.

4. Computing of an approximation of the number of solutions of the whole instance by means of the equation (25).

6 Experimental results

Consider the following results (Table 1 and Table 2) where, for reasons of hardness of the SAT problem, P/N has been taken equal respectively to 0.7 in case of 2SAT and to 2 in case of 3SAT. On the other hand, note that ap1b (resp. ap2b) concerns the cutting giving the best approximation by means of equation (24) (resp. (25)). Otherwise, ap1 and ap2 concern respectively the approximations given by (24) and (25) and obtained from the cutting detected by (27). The α parameter has been adjusted by taking into account the accuracy of the computing. Globally the new results improve the previous ones. This is more clear and significant in the case of the best cutting (see the results for ap2b with respect to those for ap1b). Even in case where the cutting is automatically obtained by means of the criterion (27), Table 1 shows some tendency of a better behavior of ap2 with respect to ap1.

This cannot be neglected if we take into account all the difficulty of the problem related to the random generation model of the SAT instances. However, it is of importance to notice that a criterion such (27) has a great capability to detect independent blocks in case of a statistical dependency hidden structure. Under these conditions and in order to improve our results we have to avoid the importance of the role of

		2SAT		3SAT	
	N	70	80	40	50
X	α	10^{-5}	10^{-6}	10^{-4}	10^{-5}
ap1b	$\frac{X}{NBS(F)} \leq \frac{1}{m}$	0.0	0.1	2.0	2.8
ap2b		0.3	0.5	2.0	2.8
ap1		4.8	5.1	12.7	13.6
ap2		4.8	5.1	13.1	14.1
ap1b	$\frac{1}{m} < \frac{X}{NBS(F)} < m$	95.8	95.2	90.7	89.0
ap2b		96.4	95.1	91.3	88.8
ap1		78.5	77.9	65.2	61.4
ap2		78.4	78.1	65.2	60.9
ap1b	$m \leq \frac{X}{NBS(F)}$	4.2	4.7	7.3	8.2
ap2b		3.3	4.4	6.7	8.4
ap1		16.7	17.0	22.1	25.0
ap2		16.8	16.8	21.7	25.0

Table 2. Percentage of instances for which the ratio $X/NBS(F)$ is limited by the boundaries $1/m$ and m ($m = 2$ for 2SAT and $m = 1.5$ for 3SAT).

the blind cutting of the seriation by means of a criterion such (27). As a matter of fact, till now, we have proposed to approximate the exact coefficient (18) by means of a coefficient such (19) having a polynomial cost and preserving the formal properties of (18) (see theorems 3 and 4). But notice that the denominator of $\mathrm{dep}(G_c, H_c)$, namely $\mathrm{P}[G_c] \times \mathrm{P}[H_c]$, is known (see point 3 of the experimental design described above in section 5). Thus, a new idea consists of evaluating $\mathrm{P}[G_c \cap H_c]$ by means of its mathematical expectation under the generation random model of the concerned SAT instance.

Indeed $G_c \cap H_c$ is union of $c \times (P - c)$ pinpoint cylinders. Some of them can be empty. In the case of rSAT, the volume of a non empty pinpoint cylinder can be $2^{N-r-j+1}$, $1 \leq j \leq r+1$. In these conditions, the mathematical expectation of the random variable associated with $\mathrm{P}[G_c \cap H_c]$, conditioned by the structure

$$\{(2^{N-r-j+1}, Q_j) \quad | \quad 1 \leq j \leq r+1\} \tag{28}$$

is given by

$$1 - \prod_{1 \leq j \leq r+1} \left(1 - 2^{(r+j-1)Q_j}\right) \tag{29}$$

(Simon and Dubois (1989), more directly Lerman (1992) cited in Lerman (1995)). By exploiting this result we have the following:

Theorem 6. *The mathematical expectation of the random variable associated with $NBS(F)$, knowing $NBS(I)$, $NBS(J)$ and (28), is given by*

$$NBS(F) = NBS(I) + NBS(J) - 2^N \prod_{1 \leq j \leq r+1} \left(1 - 2^{(r+j-1)Q_j}\right). \tag{30}$$

(30) is equal to the following formula that is in a nearest expression of (25):

$$\frac{\mathrm{NBS}(I)\times\mathrm{NBS}(J)}{2^N}+2^N\Big(\big(1-\prod_{1\leq j\leq r+1}(1-2^{(r+j-1)Q_j})\big)-\big(1-\frac{\mathrm{NBS}(I)}{2^N}\big)\big(1-\frac{\mathrm{NBS}(J)}{2^N}\big)\Big).$$

This new approximation formula will be experimented in near future.

7 Conclusion

Implicitly, we have shown in this paper the importance of the role of combinatorial data analysis in the field of computational complexity. More particularly, the problem of approximating #SAT is considered in terms of filling the logical cube $\{0,1\}^N$ by pinpoint cylinders. A proposed method (Rouat (1999), Lerman and Rouat (1999)) based on cutting seriation is more deeply studied and new results are obtained. These results are very competitive with respect those published in the literature (Bailleux and Chabrier (1996)). New and improving results can be expected following the last proposition (see theorem 6). The next stage of this research consists of introducing statistical sampling theory (Karp and Luby (1983)) in order to evaluate more accurately the dependence degree between two sets of clauses.

References

ANDRÉ, P. and DUBOIS, O. (1992): Utilisation de l'espérance du nombre de solutions afin d'optimiser la résolution du problème SAT. *C.R. de l'Académie des Sciences, Paris, 315*, 217–220.

BAILLEUX, O. and CHABRIER, J.J. (1996): Approximate resolution of hard numbering problems. *In : AAAI Thirteenth National Conference on Artificial Intelligence*, 169–174.

COOK, S.A. (1971): The complexity of theorem-proving procedures. *In : 3rd Annual ACM Symposium on the Theory of Computing*, 151–158.

DAVIS, M. and PUTNAM, H. (1960): A computing procedure for quantification theory. *Journal of the ACM, vol. 7, 3*, 201–215.

DUBOIS, O. (1991): Counting the number of solutions for instances of satisfiability. *Theoretical Computer Science, 81*, 49–64.

KARP, R.M. and LUBY, M. (1983): Monte-carlo algorithms for enumeration and reliability problems. *In : 24th IEEE Symposium of Foundations of Computer Science*, 56–64.

LEREDDE, H. (1979): *La méthode des pôles d'attraction, la méthode des pôles d'agrégation ; deux nouvelles familles d'algorithmes en classification automatique et sériation*. PhD thesis, Université de Paris VI.

LERMAN, I.C. (1972): Analyse du phénomène de la "sériation" à partir d'un tableau d'incidence. *Mathématiques et Sciences Humaines, 38*, 39–57.

LERMAN, I.C. (1995): Statistical reduction of the satisfiability problem by means of a classification method. *Data Science and its Application, Academic Press*, 219–234.

LERMAN, I.C. and ROUAT, V. (1999): Segmentation de la sériation pour la résolution de #SAT. *Mathématiques, Informatique et Sciences Humaines, 147*, 113–134.

LOZINSKII, E.L. (1992): Counting propositional models. *Information Processing Letters, 41*, 327–332.

MARCOTORCHINO, F. (1987): Block seriation problems: a unified approach. *Applied Stochastic Models and Data Analysis, vol. 3, 2*, 73–91.

PAPADIMITRIOU, C.H. (1994): *Computational complexity*. Addison Wesley.

ROUAT, V. (1999): *Validité de l'approche classification dans la réduction statistique de la complexité de #SAT*. PhD thesis, Université de Rennes 1, 1999.

ROUAT, V. and LERMAN, I.C. (1997): Utilisation de la sériation pour une résolution approchée du problème #SAT. *In : JNPC'97, résolution pratique de problèmes NP-complets*, 55–60.

ROUAT, V. and LERMAN, I.C. (1998): Problématique de la coupure dans la résolution de #SAT par sériation. *In : JNPC'98, résolution pratique de problèmes NP-complets*, 109–114.

SIMON, J.C. and DUBOIS, O. (1989): Number of solutions of satisfiability instances — applications to knowledge bases. *International Journal of Pattern Recognition and Artificial Intelligence, vol. 3, 1*, 53–65.

TODA, S. (1989): On the computational power of PP and $\oplus$P. *In : 30th Annual Symposium on Foundations of Computer Science*, 514–519.

VALIANT, L.G. (1979): The complexity of enumeration and reliability problems. *SIAM Journal on Computing, vol. 8, 3*, 410–421.