Extracting Meningitis Knowledge by Integration of Rule Induction and Association Mining

T.B. Ho, S. Kawasaki, and D.D. Nguyen

Japan Advanced Institute of Science and Technology,
Tatsunokuchi, Ishikawa, 923-1292 Japan

1 Introduction

The meningitis dataset has been used for extracting meningitis knowledge by learning and mining methods. This paper reports the result of extracting knowledge from this dataset by a novel learning method called LUPC that integrates separate-and-conquer rule induction with association rule mining. We first briefly introduce the basic ideas of LUPC then describe experiments, extracted knowledge and the result evaluation. The extracted knowledge is concerned with factors important for diagnosis (DIAG and DIAG2), for detection of bacteria or virus (CULT_FIND and CULTURE) and for predicting prognosis (C_COURSE and COURSE).

2 LUPC: Learning Unbalanced Positive Class

Consider the rule induction problem where we focus on learning a minority target class seen as the positive class C^+, denoted by *Pos*, and all other classes as the negative class C^-, denoted by *Neg*, i.e., $|Pos| << |Neg|$. Denote by cov(R) the set of instances covered by a rule R that is divided into two subsets of covered instances in *Pos* and *Neg*, denoted by $\mathrm{cov}(R) = \mathrm{cov}^+(R) \cup \mathrm{cov}^-(R)$. Our task is to find a set of predictive and descriptive rules for C^+, denoted by $R+ = \{R^+_1, R^+_2, \ldots, R^+_q\}$ so that $Pos \subseteq \mathrm{cov}(R^+_1) \cup \mathrm{cov}(R^+_2) \cup \ldots \cup \mathrm{cov}(R^+_q)$ and the discovered rules are "best" in terms of high sensitivity as well positive predictive value, and low false positive rate. Given thresholds α and β for accuracy and coverage ratio, a rule R is $\alpha\beta$-strong if $\mathrm{acc}(R) \geq \alpha$ and $|\mathrm{cov}+(R)|/|D| \geq \beta$. Table 1 presents the scheme of algorithm LUPC

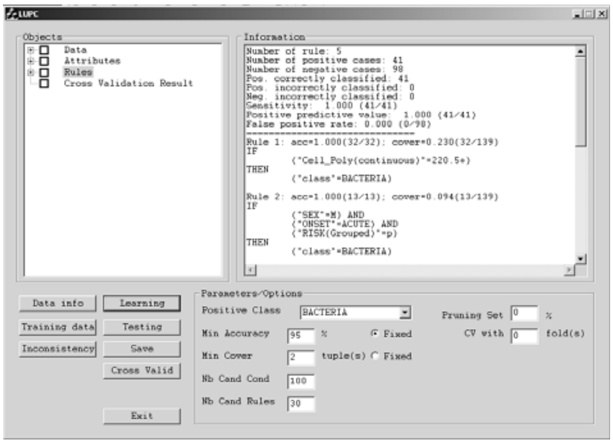
```
**Learn-positive-rule(Pos, Neg, minacc, mincov)**
RuleSet = $\emptyset$
$\alpha, \beta \leftarrow$ **Initialize**(Pos, Neg, minacc, mincov)
while (Pos $\neq \emptyset$ and $(\alpha, \beta) \neq$ (minacc, mincov))
    NewRule $\leftarrow$ **BestRule**(Pos, Neg, □□□□)
    if (NewRule $\neq \emptyset$)
        Pos $\leftarrow$ Pos ¥ $Cover^+$(NewRule)
        RuleSet $\leftarrow$ RuleSet $\cup$ NewRule
    else **Reduce**$(\alpha, \beta)$
RuleSet $\leftarrow$ PostProcess(RuleSet)
return(RuleSet)

**Procedure BestRulePos, Neg, $\alpha, \beta$)**
CandidateRuleSet = $\emptyset$
AttributeValuePairs((Pos, Neg, $\alpha, \beta$)
while StopCondition(Pos, Neg, $\alpha, \beta$)
    CandidateRules(Pos, Neg, $\alpha, \beta$)
 BestRule $\leftarrow$ First CandidateRule in
   CandidateRuleSet
return(BestRule)
```

Table 1. The scheme of algorithm LUPC

for solving effectively the above problem. There are three essential features of LUPC that make it possible to learn efficiently minority classes in unbalanced datasets. Firstly, it carries out a search biasing alternatively on accuracy and cover ratio with adaptive thresholds. Secondly, it focuses on doing separate-send-conquer induction in the target class with exploitation of the unbalanced property of datasets that allows trying the beam search with a large beam search parameter and one-sided selection. The following property shows the necessary constraint on $cov^-(R)$ for a rule R to be $\alpha\beta$-strong in terms of $cov^+(R)$ and the accuracy threshold. It will be used to reduce time of scanning the large *Neg* in generating and selecting candidate rules for C^+: given α, a rule R is not $\alpha\beta$-strong for any arbitrary β if $cov^-(R) \geq ((1-\alpha)/\alpha) \times cov^+(R)$. Thirdly, LUPC integrates pre-pruning and post-pruning in a way that can avoid over-pruning.

3 Finding Rules from Meningitis Data

We use two methods for discretizing numerical attributes in the meningitis data: entropy-based and rough set-based methods. The entropy-based method often yields few intervals of values, and ignores many attributes (15 out of 38 attributes). The rough set-based method divides continuous attributes into more intervals of values and do not ignore any attributes. From the discretized dataset we created six derived datasets with the corresponding class attribute is from DIAG, DIAG2, CULT_FIND, CULTURE, C_COURSE and



```
Number of rule: 5
Number of positive cases: 41
Number of negative cases: 98
Pos. correctly classified: 41
Pos. incorrectly classified: 0
Neg. incorrectly classified: 0
Sensitivity:  1.000 (41/41)
Positive predictive value:  1.000 (41/41)
False positive rate: 0.000 (0/98)
--------------------------------
Rule 1: acc=1.000(32/32); cover=0.230(32/139)
IF
        ("Cell_Poly(continuous)"=220.5+)
THEN
        ("class"=BACTERIA)

Rule 2: acc=1.000(13/13); cover=0.094(13/139)
IF
        ("SEX"=M) AND
        ("ONSET"=ACUTE) AND
        ("RISK(Grouped)"=p)
THEN
        ("class"=BACTERIA)
```

Parameters/Options

Positive Class: BACTERIA
Pruning Set: 0 %
Min Accuracy: 95 % ● Fixed
CV with: 0 fold(s)
Min Cover: 2 tuple(s) ○ Fixed
Nb Cand Cond: 100
Nb Cand Rules: 30

Figure 1. Finding meningitis knowledge with LUPC

COURSE. We run LUPC on each of these datasets on two modes: learning one target class and learning all classes. Experiments have been done with fixed default parameters for finding rules: 95% for minimum accuracy of a rule, 2 cases are minimum cover of a rule, 100 and 30 are numbers of candidate attribute-value pairs and rules, respectively. Different rules were extracted and they are synthesized in nearly 80 tables in the Excel format according to the derived datasets and learning modes, for example:

IF LOC = [*-1) and
ONSET = ACUTE and
CSF_CELL = [1505-*) and
CELL_POLY = [431-*)
THEN class = BACTERIA [accuracy = 1.00 (12/12); cover = 0.086]

Based on synthesized tables of discovered rules, we have provided the domain experts a number of observations and analysis that are commonly concerned with the most frequent attributes in each class, the significant attributes or attribute-value pairs, the significant co-occurred attribute-values pairs, the strong rules with particularly large coverage if available, and rules that may be exceptional.

Factors Important for Meningitis Diagnosis DIAG and DIAG2

From discovered rules for DIAG we observed that:

- most frequent attributes: Cell_Poly, Loc_Dat, Egg_Focus, Focal, Ct_Find.

- significant attributes or attribute-value pairs:
  - "Cell_Poly > 220.5" for BACTE(E) and BACTERIA,
  - "Cell_Poly < 220.5" for VIRUS and VIRUS(E),
  - "Egg_Focus = +" for VIRUS(E),
  - "Ct_find = abnormal" for ABSCESS.
- significant co-occurred attribute-values pairs:
  - "Cell_Poly < 220.5" AND "Egg_Focus = -" for VIRUS,
  - "Cell_Poly < 220.5" AND "Focal = +" for VIRUS(E).

And from discovered rule for DIAG2:

- most frequent attributes: Focal, Cell_Poly, Loc_Data, Egg_Focus, Ct_Find.
- significant or discriminant attributes or attribute-value pairs are reconfirmed
  - "Ct_find = abnormal" for ABSCESS,
  - "Cell_Poly geq 220.5" for BACTE(E) and BACTERIA,
  - "Cell_Poly < 220.5" for VIRUS and VIRUS(E).
- significant co-occurred attribute-values pairs: reconfirmed the above conclusions and some new as "Cell_Poly > 220.5" AND "Onset = Acute" AND "Loc = -1.5" for BACTERIA.
- rules with large coverage: rules for VIRUS
- rules that may be special or typical: rule 1 for ABSCESS, rule 2 for BACTERIA.

A general observation is there are big groups of VIRUS cases that share common symptoms (VIRUS rules with bigger coverage but not very high accuracy) while the rules for BACTERIA are with relatively smaller coverage but higher accuracy. The attribute "ONSET" has high frequency but seems not significant in distinguishing diseases.

Factors for Predicting Prognosis C_COURSE and COURSE

From discovered rules for C_COURSE we observed that:

- most frequent attributes: Lasegue, Focal, Loc_Dat, Onset Ct_Find.
- significant or discriminant attributes or attribute-value pairs:
  - for class "dead": Locdat = +", "Egg_wave = abnormal",
  - for class "negative": "Onset = Acute", "Lasegue = 0", "Focal = -", "Cell_Mono > 10".
- significant co-occurred attribute-values pairs:
  - "Cell_Mono < 10" AND "Locdat = +" for class "dead",

  - “Egg_wave = abnormal” AND “Locdat = +” for class “dead”,
  - “Kernig = 0” AND “Focal = -” AND “Crp < 4.8” for class “negative”,
  - “Kernig = 0” AND “Focal = -” AND “Csf_Cell in (30.5-1040)” for class “negative”.
- rules with large coverage: rules from 5 to 17 for class “negative”.
- rules that may be special or typical: all rules for class “dead”, rule 23 for class “negative”.

And form rules for COURSE:

- most frequent attributes: Lasegue, Focal, Locdat.
- significant or discriminant attributes or attribute-value pairs:
  - “Focal = -” in class “n” and “Focal = +” in class “p”,
  - “Locdat = -” in class “n” and “Locdat = +” in class “p”,
  - “Egg_wave = normal” in class “n”, “Egg_wave = abnormal” in “p”,
  - “Cell_Mono > 10” in class “n” and “Cell_Mono < 10” in class “p”
  - “Lasegue = 0” is popular in class “n”.
- significant co-occurred attribute-values pairs:
  - “Lasegue = 0” AND “Focal = -” AND “Crp < 4.8” in class “n”,
  - “Lasegue = 0” AND “Cell_Mono > 1.0” in class “n”,
  - “Local = +” AND “Focal = +” AND “Egg_wave = abnormal” in “p”,
  - “Locdat = +” AND “Cell_Mono < 1.0” in class “p”.
- rules with large coverage: most rules for class “n”.

Two classes “n” and “p” can be distinguished by obtained rules.

Detection of Bacteria or Virus: CULTURE and CULT_FIND

From discovered rules for CULTURE we observed that:

- most frequent attributes: Loc_Dat, Crp, Ct_Find, Csf_Cell.
- significant or discriminant attributes or attribute-value pairs:
  - “Locdat = -”, “Crp < 4.8”, “Cell_Mono > 10” are pupolar in class “-”,
  - “Egg_wave = abnormal”, Ct_find = abnormal” are popular in classes “he pes” and “strepto”
- significant co-occurred attribute-values pairs:
  - “Locdat = -” AND “Crp < 4.8” AND “Cell_Mono > 10” in class “-”,
  - “Egg_wave = abnormal” AND “Ct_find = abnormal” OR “Egg_wave = abnormal” AND “Risk = sinutisis” in class “strepto”.
- rules with large coverage: most rules for class “-”.

And from rules for CULTFIND:

- most frequent attributes: Loc_Dat, Egg_Focus, Csf_Cell, Cf_Find, Risk.

- significant or discriminant attributes or attribute-value pairs:
  - "Locdat = -" is popular in "F" while "Locdat = +" is popular in "T",
  - "Crp < 4.8" is popular in "F" while "Crp > 4.8" is popular in "T",
  - "Cell_Mono > 10" is popular in "F" while "Cell_Mono < 10" is popular in "T",
  - "Ct_find = normal" is popular in "F" while "Ct_find = abnormal" is popular in "T",
  - "Risk = p" is popular in "F" while "Risk = n" OR "Risk = sinusitis" are popular in "T".
- significant co-occurred attribute-values pairs:
  - "Onset = acute" AND "Crp < 4.8" in "F",
  - "LocDat = +" AND "Risk = n" in "T".

4 Conclusion

We have briefly introduced method LUPC to learn the target positive class from large unbalanced datasets. The essence of LUPC is its combination of separate-and-conquer rule induction with association rules mining, as well the use of dynamic multiple thresholds and the property of unbalanced datasets. We apply LUPC to investigate the meningitis dataset. Many rules with high accuracy have been found for factors important for diagnosis (DIAG and DIAG2), for detection of bacteria or virus (CULT_FIND and CULTURE) and for predicting prognosis (C_COURSE and COURSE). Appendixes 1 and 2 present a summarization of rules extracted for DIAG.

Literature

1. Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. International Conference Management of Data SIGMOD'93, 207-216.
2. Brunk, C. A.and Pazzani, M. J. (1991). An Investigation of Noise-Tolerant Relational Concept Learning Algorithms, Eight International Conference on Machine Learning, 389-393.
3. Furnkranz, J. (1999). Separate-and-Conquer Rule Learning. Journal Artificial Intelligence Review, **13**, 3-54.
4. Tsumoto, S. (2000). Comparison and Evaluation of Knowledge Obtained by KDD Methods. Journal of Japanese Society for Artificial Intelligence, Vol. 15, N. 5, 790-797.

Appendix 1. LUPC's rule learning includes two modes: learning all classes and learning only one target class. These four tables show the numbers of cases which coverd rules from "DIAG2" and "DIAG" obtained by LUPC with the condition of: (1) learning mode: all classes, (2) Minimum accuracy: 95%, (3) Minimum cover: 2 cases, (4) Number candidate conditions: 100, (5) Number candidate rules: 30. Table 3 is the result on "DIAG2" discretized by entropy. Table 4 is the result on "DIAG2" discretized by Rosetta. Likewise, Table 5 and Table 6 are the results on "DIAG2" discretized by entropy and Rosetta.

Table3: Rules from "diag2" discretized by entropy

class	ID	(a)	(b)	(c)	(d)
BACTERIA	1	1.0	32	32	0.2
BACTERIA	2	1.0	13	13	0.0
BACTERIA	3	1.0	12	12	0.0
BACTERIA	4	1.0	11	11	0.0
BACTERIA	5	1.0	8	8	0.0
VIRUS	6	0.9	100	95	0.7
VIRUS	7	0.9	88	85	0.6
VIRUS	8	0.9	83	82	0.6
VIRUS	7	0.9	88	85	0.6
VIRUS	8	0.9	83	82	0.6

(a) : accuracy
(b) : number of covered cases
(c) : number of correct cases
(d) : coverage of the rule

Table 4: Rules from "diag2" discretized by Rosetta

class	ID	(a)	(b)	(c)	(d)
BACTERIA	1	1.0	27	27	0.1
BACTERIA	2	1.0	15	15	0.11
BACTERIA	3	1.0	14	14	0.1
BACTERIA	4	1.0	12	12	0.0
BACTERIA	5	1.0	9	9	0.0
BACTERIA	6	1.0	9	9	0.0
BACTERIA	7	1.0	9	9	0.0
BACTERIA	8	1.0	5	5	0.0
VIRUS	9	0.9	47	45	0.3
VIRUS	10	0.9	45	43	0.3
VIRUS	11	0.9	45	43	0.3
VIRUS	12	0.9	45	43	0.3
VIRUS	13	0.9	44	42	0.3
VIRUS	14	0.9	43	41	0.3
VIRUS	15	0.9	42	40	0.3
VIRUS	16	0.9	42	40	0.3
VIRUS	17	0.9	40	38	0.2
VIRUS	18	0.9	32	31	0.2
VIRUS	19	0.9	32	31	0.2
VIRUS	20	0.9	29	28	0.2
VIRUS	21	0.9	28	27	0.2
VIRUS	22	0.9	27	26	0.1
VIRUS	23	0.9	26	25	0.1
VIRUS	24	1.0	23	23	0.1
VIRUS	25	0.9	23	22	0.1
VIRUS	26	0.9	23	22	0.1

Table 5: Rules from "diag" discretized by entropy

class	ID	(a)	(b)	(c)	(d)
ABSCESS	1	1.0	6	6	0.0
ABSCESS	2	1.0	3	3	0.0
ABSCESS	3	1.0	2	2	0.0
ABSCESS	4	1.0	2	2	0.0
BACTE(E)	5	1.0	3	3	0.0
BACTE(E)	6	1.0	2	2	0.0
BACTE(E)	7	1.0	2	2	0.0
BACTERIA	8	1.0	11	11	0.0
BACTERIA	9	1.0	10	10	0.0
BACTERIA	10	1.0	8	8	0.0
BACTERIA	11	1.0	8	8	0.0
BACTERIA	12	1.0	8	8	0.0
BACTERIA	13	1.0	6	6	0.0
BACTERIA	14	1.0	5	5	0.0
VIRUS	15	0.9	61	58	0.4
VIRUS	16	0.9	60	57	0.4
VIRUS	17	0.9	58	56	0.4
VIRUS	18	0.9	54	52	0.3
VIRUS	20	0.9	43	41	0.3
VIRUS	19	0.9	51	49	0.3
VIRUS(E)	21	1.0	11	11	0.0
VIRUS(E)	22	1.0	11	11	0.0
VIRUS(E)	23	1.0	10	10	0.0
VIRUS(E)	24	1.0	9	9	0.0
VIRUS(E)	25	1.0	9	9	0.0
VIRUS(E)	26	1.0	8	8	0.0
VIRUS(E)	27	1.0	6	6	0.0
VIRUS(E)	28	1.0	7	7	0.0

Table6: Rules from "diag" discretized by Rosetta

class	ID	(a)	(b)	(c)	(d)
ABSCESS	1	1.0	6	6	0.0
ABSCESS	2	1.0	4	4	0.0
ABSCESS	3	1.0	4	4	0.0
ABSCESS	4	1.0	4	4	0.0
BACTE(E)	5	1.0	3	3	0.0
BACTE(E)	6	1.0	3	3	0.0
BACTE(E)	7	1.0	3	3	0.0
BACTE(E)	8	1.0	3	3	0.0
BACTE(E)	9	1.0	2	2	0.0
BACTE(E)	10	1.0	2	2	0.0
BACTERIA	11	1.0	12	12	0.0
BACTERIA	12	1.0	9	9	0.0
BACTERIA	13	1.0	8	8	0.0
BACTERIA	14	1.0	7	7	0.0
BACTERIA	15	1.0	7	7	0.0
BACTERIA	16	1.0	6	6	0.0
BACTERIA	17	1.0	6	6	0.0
BACTERIA	18	1.0	6	6	0.0
BACTERIA	19	1.0	5	5	0.0
BACTERIA	20	1.0	5	5	0.0
BACTERIA	21	1.0	4	4	0.0
VIRUS	22	0.9	22	21	0.1
VIRUS	23	0.9	21	20	0.1
VIRUS	24	0.9	21	20	0.1
VIRUS	25	0.9	21	20	0.1
VIRUS	26	0.9	20	19	0.1
VIRUS	27	1.0	18	18	0.1
VIRUS	28	1.0	15	15	0.11
VIRUS	29	1.0	15	15	0.11
VIRUS	30	1.0	14	14	0.1
VIRUS	31	1.0	14	14	0.1
VIRUS	32	0.9	33	32	0.2
VIRUS(E)	33	1.0	10	10	0.0
VIRUS(E)	34	1.0	10	10	0.0
VIRUS(E)	35	1.0	9	9	0.0
VIRUS(E)	36	1.0	9	9	0.0
VIRUS(E)	37	1.0	9	9	0.0
VIRUS(E)	38	1.0	7	7	0.0
VIRUS(E)	39	1.0	7	7	0.0
VIRUS(E)	40	1.0	7	7	0.0
VIRUS(E)	41	1.0	7	7	0.0
VIRUS(E)	42	1.0	7	7	0.0
VIRUS(E)	43	1.0	6	6	0.0
VIRUS(E)	44	1.0	6	6	0.0
VIRUS(E)	45	1.0	6	6	0.0

Appendix 2. Rules from “diag” with Rosetta discretization for all classes

						Values of attributes contained in each rule																															
class	rule_ID	accuracy	# of covered	# of corrct cases	coverage	AGE	SEX	COLD	HEADACHE	FEVER	NAUSEA	LOC	SEIZURE	ONSET	BT	STIFF	KERNIG	LASEGUE	GCS	LOC_DAT	FOCAL	WBC	CRP	ESR	CT_FIND	EEG_WAVE	EEG_FOCUS	CSF_CELL	Cell_Poly	Cell_Mono	CSF_PRO	CSF_GLU	CULT_FIND	CULTURE	CSF_CELL7	C_COURSE	COURSE
ABSCESS	1	1.00	6	6	0.04			[*-1)						ACUTE											abnor-			[*-75)									
ABSCESS	2	1.00	4	4	0.03															+		[11550-*)													[*-1)		
ABSCESS	3	1.00	4	4	0.03								[*-1)						[14-*)						abnor-							[65-*)			[*-1)		
ABSCESS	4	1.00	4	4	0.03			[*-1)												-					abnor-		-										n
BACTE(E)	5	1.00	3	3	0.02																							[1505-*)				[65-*)					
BACTE(E)	6	1.00	3	3	0.02	[52-*)						[1-2)																[1505-*)									
BACTE(E)	7	1.00	3	3	0.02					[1-4)																abnormal			[431-*)								
BACTE(E)	8	1.00	3	3	0.02			[*-1)												+	-						-		[431-*)								
BACTE(E)	9	1.00	2	2	0.01																								[431-*)		[35-68)						
BACTE(E)	10	1.00	2	2	0.01																		[2.1-4.6)								[35-68)						
BACTERIA	11	1.00	12	12	0.09							[*-1)		ACUTE														[1505-*)	[431-*)								
BACTERIA	12	1.00	9	9	0.06						[*-1)	[*-1)		ACUTE															[431-*)								
BACTERIA	13	1.00	8	8	0.06																-								[431-*)	[750-*)							
BACTERIA	14	1.00	7	7	0.05																								[431-*)			[*-44)					
BACTERIA	15	1.00	7	7	0.05			[*-1)						ACUTE					[14-*)								-		[431-*)				F				
BACTERIA	16	1.00	6	6	0.04									ACUTE						-			[4.6-*)														
BACTERIA	17	1.00	6	6	0.04																								[431-*)						[395-*)		
BACTERIA	18	1.00	6	6	0.04									ACUTE											normal		-										
BACTERIA	19	1.00	5	5	0.04																-													strepto			
BACTERIA	20	1.00	5	5	0.04																										[68-91)						
BACTERIA	21	1.00	4	4	0.03		M																							[47-125)			T				
VIRUS	22	0.95	22	21	0.16																							[75-219)						-		negative	
VIRUS	23	0.95	21	20	0.15				[3-5)																normal												
VIRUS	24	0.95	21	20	0.15										[37.3-38.1)						-				normal												
VIRUS	25	0.95	21	20	0.15																-										[35-68)					negative	
VIRUS	26	0.95	20	19	0.14																-										[35-68)		F				
VIRUS	27	1.00	18	18	0.13									ACUTE											normal				[43-431)								
VIRUS	28	1.00	15	15	0.11	[24-31)														-	-						-										
VIRUS	29	1.00	15	15	0.11															-	-									[125-327)						negative	
VIRUS	30	1.00	14	14	0.10															-				[*-1)			-		[5-18)								
VIRUS	31	1.00	14	14	0.10	[24-31)														-	-																
VIRUS	32	1.00	33	32	0.24									ACUTE						-	-		[*-0.2)				-										
VIRUS(E)	33	1.00	10	10	0.07		F									[2-3)											+										
VIRUS(E)	34	1.00	10	10	0.07		F																[*-0.2)				+										
VIRUS(E)	35	1.00	9	9	0.06		F													+	+																
VIRUS(E)	36	1.00	9	9	0.06		F																														p
VIRUS(E)	37	1.00	9	9	0.06		F														+			[*-1)													
VIRUS(E)	38	1.00	7	7	0.05															+									[5-18)								
VIRUS(E)	39	1.00	7	7	0.05															+									[18-43)								
VIRUS(E)	40	1.00	7	7	0.05						[*-1)					[2-3)												[501-1505)					T				
VIRUS(E)	41	1.00	7	7	0.05																						+				[91-133)						
VIRUS(E)	42	1.00	7	7	0.05									ACUTE							+						+										
VIRUS(E)	43	1.00	6	6	0.04																	[5950-6150)											F				p
VIRUS(E)	44	1.00	6	6	0.04								[*-1)									[5950-6150)							[*-5)								
VIRUS(E)	45	1.00	6	6	0.04						[*-1)										+		[*-0.2)		normal												