Evaluating Hypothesis and Experimental Design - #2

Patricia J Riddle

Computer Science 760

Four Spurious Effects

Ceiling Effect - Holte's 1R example

Data set	IR	LA	LY	MU	SE	SO	VO	VI
C4	93.8	77.2	77.5	100	97.7	97.5	95.6	89.4
1R	95.9	87.4	77.3	98.4	95	87	95.2	87.9
Max	95.9	87.4	77.5	100	97.7	97.5	95.6	89.4

Other 3 Effects

Regression Effects - if chance plays a role, then always run the same problems

Order Effects - counter balancing or at least a few orders

Sampling Bias - how data was collected is very important - the independent variable can change the location of the distribution but not its shape

Experiments with Standard Deviation

ex	name	C4.5 P	C4.5 error rate	Randomized C4.5 P	Randomized C4.5 error rate	Bagged C4.5 P	Bagged C4.5 error rate	Adaboosted C4.5 P	Adaboosted C4.5 error rate
	sonar		0.3257±0.0637		0.2018±0.0545	*	0.2752±0.0607	*	0.1651±0.0505
	letter		0.1225±0.0045		0.0285±0.0023		0.0552±0.0032	*	0.0271±0.0023
	splice	*	0.0575±0.0081	*	0.0397±0.0068	*	0.0506±0.0076		0.0503±0.0076
	segment		0.0328±0.0073		0.0203±0.0058		0.0263±0.0065		0.0151±0.0050
	glass	*	0.3437±0.0636		0.2277±0.0562		0.2723±0.0596	*	0.2277±0.0562
	soybean		0.1262±0.0371	*	0.0852±0.0312	*	0.1009±0.0337	*	0.0757±0.0296
	autos		0.2326±0.0578	*	0.1581±0.0499		0.1814±0.0528		0.1814±0.0528
	satimage	*	0.1515±0.0157		0.0890±0.0125		0.1020±0.0133		0.0850±0.0122
	annealing	*	0.0132±0.0075		0.0088±0.0061		0.0099±0.0065		0.0055±0.0048
	krk		0.1887±0.0046		0.1309±0.0039		0.1463±0.0041	*	0.1026±0.0036
	heart-v	*	0.2762±0.0620	*	0.2429±0.0594		0.2619±0.0609	*	0.2810±0.0623
	heart-c	*	0.2396±0.0481	*	0.1853±0.0437	*	0.1981±0.0449		0.2045±0.0454
	breast-y	*	0.2601±0.0508	*	0.2500±0.0502	*	0.2635±0.0511	*	0.3142±0.0538
	phoneme	*	0.1661±0.0086		0.1437±0.0081		0.1509±0.0082	*	0.1464±0.0081
	voting	*	0.1146±0.0299	*	0.0921±0.0272	*	0.0966±0.0278	*	0.1034±0.0286
	vehicle		0.2944±0.0307		0.2477±0.0291		0.2570±0.0294		0.2196±0.0279
	lymph		0.1962±0.0640		0.1772±0.0615		0.1835±0.0624	*	0.1266±0.0536
	breast-w	*	0.0494±0.0161	*	0.0353±0.0137		0.0367±0.0139		0.0310±0.0128
	credit-g	*	0.2921±0.0282		0.2416±0.0265	*	0.2495±0.0268		0.2347±0.0263
	primary	*	0.5845±0.0525	*	0.5501±0.0530		0.5645±0.0528	*	0.5960±0.0522
	shuttle		0.0003±0.0003		0.0002±0.0002		0.0002±0.0002		0.0001±0.0002
	heart-s	*	0.0677±0.0444	*	0.0677±0.0444	*	0.0677±0.0444	*	0.0902±0.0506
	iris		0.0563±0.0369	*	0.0500±0.0349	*	0.0500±0.0349	*	0.0688±0.0405
	sick	*	0.0132±0.0036		0.0137±0.0037		0.0137±0.0037	*	0.0095±0.0031
	hepatitis		0.1758±0.0599		0.1636±0.0582		0.1636±0.0582	*	0.1636±0.0582
	credit-a	*	0.1614±0.0275	*	0.1400±0.0259		0.1371±0.0257	*	0.1300±0.0251
	waveform	*	0.2341±0.0117		0.1784±0.0106		0.1675±0.0104		0.1521±0.0100
	horse-colic	*	0.1561±0.0371		0.1561±0.0371		0.1481±0.0363	*	0.1825±0.0395
	heart-h	*	0.1645±0.0424	*	0.1809±0.0440	*	0.1579±0.0417		0.2039±0.0461
	labor		0.1493±0.0925	*	0.1493±0.0925		0.1194±0.0842	*	0.1194±0.0842
	krkp		0.0075±0.0030		0.0075±0.0030		0.0056±0.0026	*	0.0037±0.0021
	audiology		0.2203±0.0540	*	0.2458±0.0561		0.1822±0.0503	*	0.1525±0.0469
	hypo		0.0058±0.0024	*	0.0079±0.0028		0.0042±0.0021	*	0.0040±0.0020

Types of Error

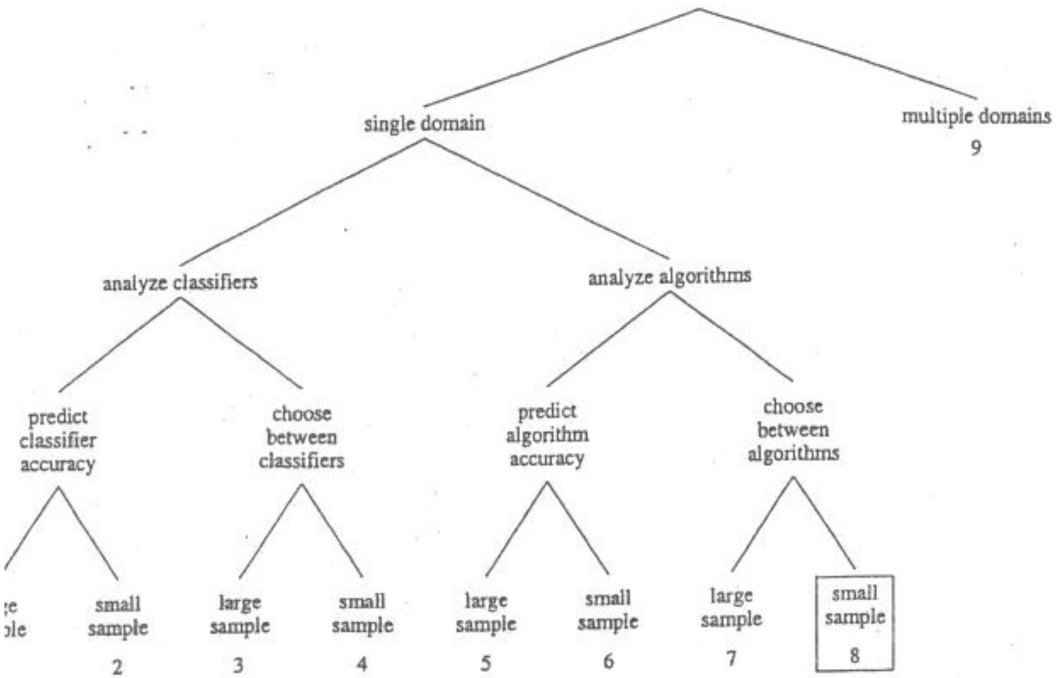
Type I error, also known as an "error of the first kind", an α error, or a "false positive": the error of **rejecting a null hypothesis when it is actually true**.

(we thought they were statistically significantly different and they were the same)

Type II error, also known as an "error of the second kind", a β error, or a "false negative": the error of **failing to reject a null hypothesis when it is in fact not true**.

(we thought they were the same and they were statistically significantly different)

Statistical Questions in Machine Learning



ure 1: A taxonomy of statistical questions in machine learning. The boxed node (Question 8) ıe subject of this paper.

What is the question?

- Classifying Unseen Examples?
- Learning New Classifiers in Future?

Examples

- Driving a Car?
- Learning a New Car Driver in Future?
- Teach someone to Drive a car?

Question Assumptions

We assume that all datapoints (examples) are drawn independently from a fixed probability distribution defined by the particular problem.

Independently?

This is almost never the case!!!

Comparison of Two Classifiers (with lots of datasets)

- Wilcoxon Signed-Ranks Test
  - Non-parametric alternative to the paired t-test

Notation

- $\mathbf{d_i}$ - the difference between the performance scores of the two classifiers on **i**-th out of **N** data sets.

- The differences are ranked according to their absolute values; average ranks are assigned in case of ties.

- Let **R+** be the sum of ranks for the data sets on which the second algorithm outperformed the first, and **R−** the sum of ranks for the opposite.

- Ranks of $\mathbf{d_i} = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

Formulas

$$R^{+} = \sum_{d_i > 0} rank(d_i) + \frac{1}{2} \sum_{d_i = 0} rank(d_i)$$

$$R^{-} = \sum_{d_i < 0} rank(d_i) + \frac{1}{2} \sum_{d_i = 0} rank(d_i)$$

Formulas

- Let **T** be the smaller of the sums, **T** = min(**R+**,**R−**).
- Most books on general statistics include a table of exact critical values for **T** for **N** up to 25 (or sometimes more).
- For a larger number of data sets, the statistics

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}$$

- is distributed approximately normally.
- With $\alpha = 0.05$, the null-hypothesis can be rejected if **z** is smaller than -1.96.

Why Wilcoxon is better than Paired T-test

- It assumes commensurability of differences, but only qualitatively: greater differences still count more, which is probably desired, but the absolute magnitudes are ignored.

- From the statistical point of view, the test is safer since it does not assume normal distributions.

- Also, the outliers (exceptionally good/bad performances on a few data sets) have less effect on the Wilcoxon than on the t-test.

Wilcoxon Usage Tips

- The Wilcoxon test assumes continuous differences $\mathbf{d_i}$, therefore they should not be rounded to, say, one or two decimals since this would decrease the power of the test due to a high number of ties.
- When the assumptions of the paired t-test are met, the Wilcoxon signed-ranks test is less powerful than the paired t-test.
- On the other hand, when the assumptions are violated, the Wilcoxon test can be even more powerful than the t-test.

Presentation of Results

- A popular way to compare the overall performances of classifiers is to count the number of data sets on which an algorithm is the overall winner.

- When multiple algorithms are compared, pairwise comparisons are sometimes organized in a matrix.

Experiments with Pairwise Combination Chart

Table 3. All pairwise combinations of the four methods for four levels of noise and 9 domains. Each cell contains the number of wins, losses, and ties between the algorithm in that row and the algorithm in that column.

Noise ≈ 0%	C4.5	Adaboost C4.5	Bagged C4.5
Random C4.5	5 – 0 – 4	1 – 6 – 2	3 – 3 – 3
Bagged C4.5	4 – 0 – 5	0 – 5 – 4	
Adaboost C4.5	6 – 0 – 3		

Noise ≈ 5%	C4.5	Adaboost C4.5	Bagged C4.5
Random C4.5	5 – 2 – 2	3 – 2 – 4	1 – 5 – 3
Bagged C4.5	6 – 0 – 3	5 – 1 – 3	
Adaboost C4.5	3 – 3 – 3		

Noise ≈ 10%	C4.5	Adaboost C4.5	Bagged C4.5
Random C4.5	4 – 1 – 4	5 – 1 – 3	1 – 6 – 2
Bagged C4.5	5 – 0 – 4	6 – 1 – 2	
Adaboost C4.5	2 – 3 – 4		

Noise ≈ 20%	C4.5	Adaboost C4.5	Bagged C4.5
Random C4.5	5 – 2 – 2	5 – 0 – 4	0 – 2 – 7
Bagged C4.5	7 – 0 – 2	6 – 0 – 3	
Adaboost C4.5	3 – 6 – 0		

Comparing Wins and Losses

- Since tied matches support the null-hypothesis we should not discount them but split them evenly between the two classifiers; if there is an odd number of them, we again ignore one.

- Some authors prefer to count only the significant wins and losses, where the significance is determined using a statistical test on each data set, for instance Dietterich's 5**x**2cv. The reasoning behind this practice is that "some wins and losses are random and these should not count".

- This would be a valid argument if statistical tests could distinguish between the random and non-random differences. However, statistical tests only measure the improbability of the obtained experimental result if the null hypothesis was correct, which is not even the (im)probability of the null-hypothesis.

Comparing Multiple Classifiers (with lots of datasets)

- The Friedman test (Friedman, 1937, 1940) is a non-parametric equivalent of the repeated-measures ANOVA.
- It ranks the algorithms for each data set separately, the best performing algorithm getting the rank of 1, the second best rank 2. . . .
- In case of ties, average ranks are assigned.

Friedman test **statistic**

- Let r^j_i be the rank of the j-th of k algorithms on the i-th of N data sets.

- The Friedman test compares the average ranks of algorithms, $R_j = (1/N)\ \boldsymbol{\Sigma_i r^j_i}$

The statistic

- Under the null-hypothesis, which states that all the algorithms are equivalent and so their ranks R_j should be equal, the Friedman statistic

$$X_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

- is distributed according to χ^2_F with k−1 degrees of freedom, when N and k are big enough (as a rule of a thumb, $N > 10$ and $k > 5$).
- For a smaller number of algorithms and data sets, exact critical values have been computed (Zar, 1998; Sheskin, 2000).

The **Iman-Davenport** T2 variant of the Friedman test **statistic**

- Iman and Davenport (1980) showed that Friedman's χ^2_F is undesirably conservative and derived a better statistic

$$F_F = \frac{(N-1)\mathrm{X}_F^2}{N(k-1)-X_F^2}$$

- which is distributed according to the F-distribution with k −1 and (k−1)(N−1) degrees of freedom.
- The table of critical values can be found in any statistical book.

Nemenyi test

- If the null-hypothesis is rejected, we can proceed with a post-hoc test. The Nemenyi test (Nemenyi,1963) is similar to the Tukey test for ANOVA and is used when all classifiers are compared to each other.

- The performance of two classifiers is significantly different if the corresponding average ranks differ by at least the critical difference

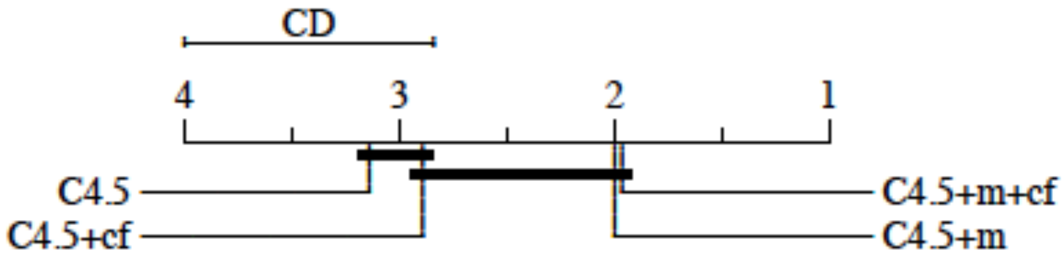
$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

- where critical values $\mathbf{q}_a$ are based on the Studentized range statistic divided by $\sqrt{2}$.

Holm Test

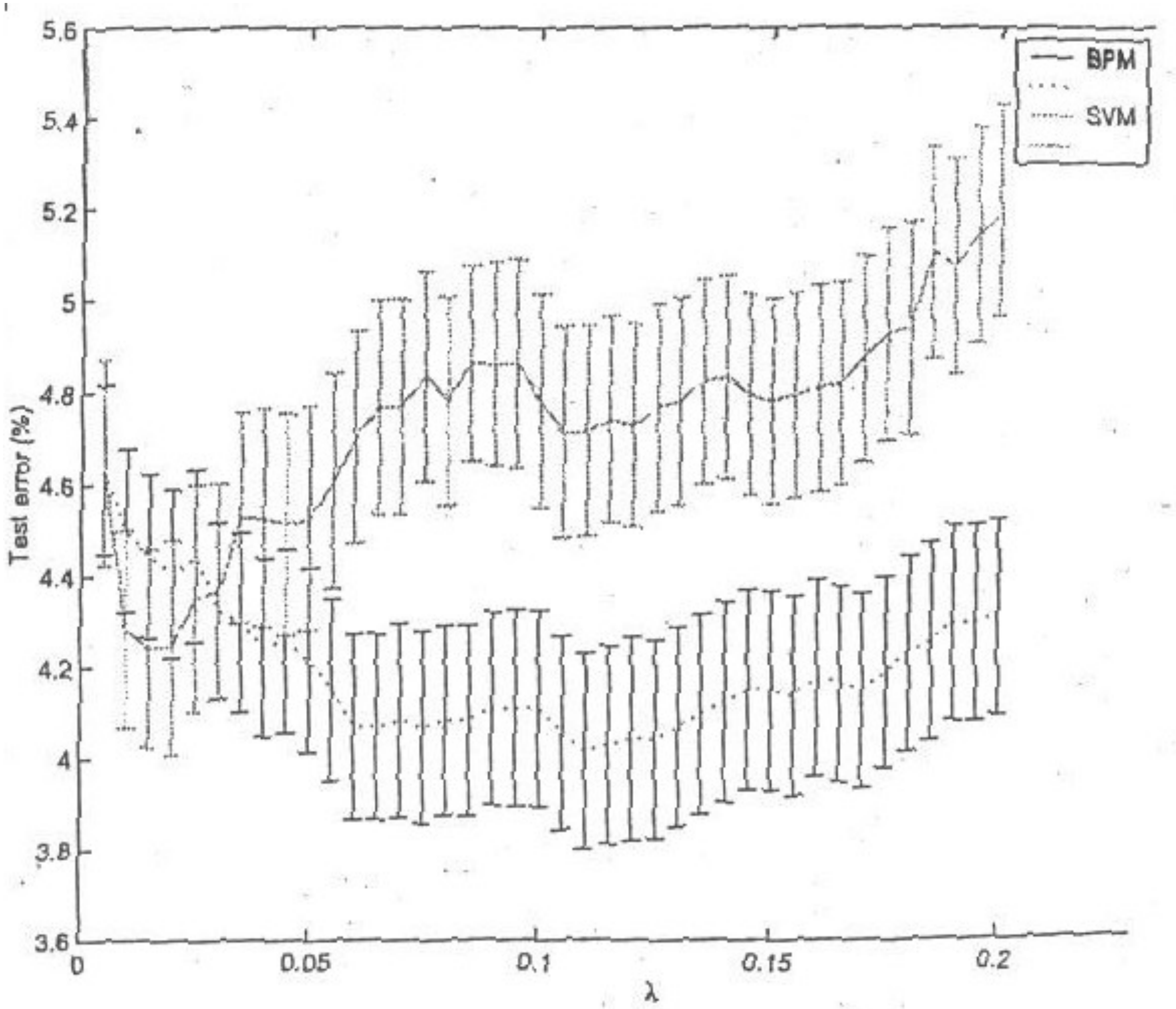
- We will denote the ordered p values by p_1, p_2, ..., so that $p_1 \leq p_2 \leq \ldots p_{k-1}$. The simplest such methods are due to Holm (1979) and Hochberg (1988).
- They both compare each p_i with $\alpha/(k-i)$, but differ in the order of the tests.
- Holm's step-down procedure starts with the most significant p value.
- If p_1 is below $\alpha/(k-1)$, the corresponding hypothesis is rejected and we are allowed to compare p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, the test proceeds with the third, and so on.
- As soon as a certain null hypothesis cannot be rejected, all the remaining hypotheses are retained as well.

Critical Difference Graph

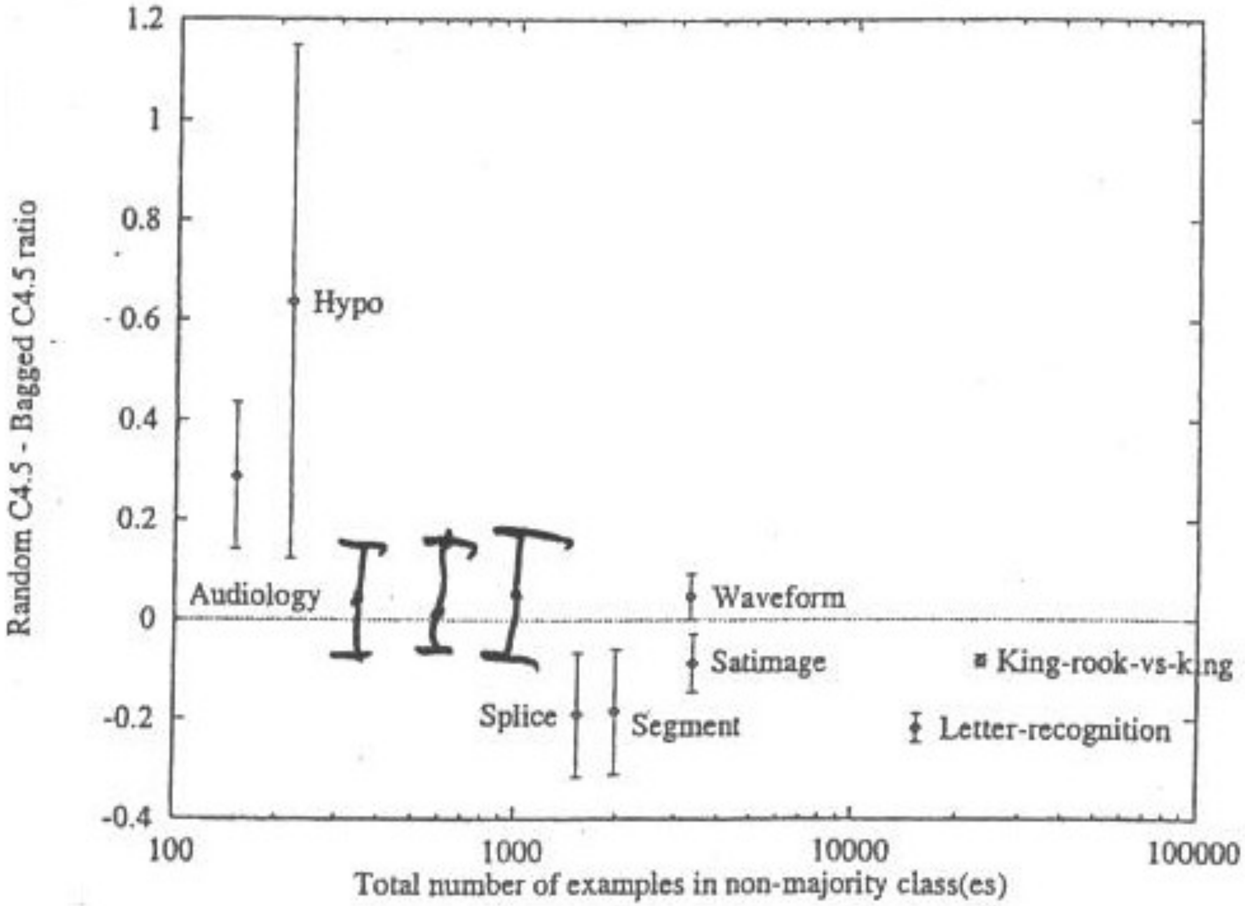


(a) Comparison of all classifiers against each other with the Nemenyi test. Groups of classifiers that are not significantly different (at $p = 0.10$) are connected.

Experiments with Learning Curves



Experiments with Difference in Performance Graph



Summary

What questions are we interested in asking?

Wilcoxon and Friedmans Test

Problems to watch out for in experimental design

Real cause of overfitting.

References

- Statistical Comparisons of Classifiersover Multiple Data Sets, Janez Demšar
- Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, Dietterich, T. G

Questions you should be able to answer?

- What is the difference between type I and type II error?
- Which of these do we worry about the most and why?
- What is the main problem with a paired t-test?

References

- *Experimental Methods for Artificial Intelligence*, Paul R. Cohen, 1995
- https://mitpress.mit.edu/books/empirical-methods-artificial-intelligence

- Multiple Comparisons in Induction Algorithms, DAVID D. JENSEN jensen@cs.umass.edu, PAUL R. COHEN cohen@cs.umass.edu
- https://link.springer.com/content/pdf/10.1023%2FA%3A1007631014630.pdf

- Statistical Comparisons of Classifiers over Multiple Data Sets, Janez Demsar, JANEZ.DEMSAR@FRI.UNI-LJ.SI
- http://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf