



· 三篇论文 ·

SCIENCE AND TECHNOLOGY

《The Missing Links: Bugs and Bug-fix Commits》	2010	FSE/ESEC
---	------	----------

《Is better data better than better data miners?: on the benefits of tuning smote for defect prediction》	2018	ICSE
---	------	------



The Missing Links: Bugs and Bug-fix Commits

2010 FSE/ESEC

Given the wide use of linked defect data, it is vital to gauge the nature and extent of the bias, and try to develop testable theories and models of the bias.

Contribution

Link



Issues

Bug-feature Bias

where only the fixes of certain types of defects are linked

Commit-feature Bias

where only the certain kinds of fixes, or fixes to certain kinds of files, are linked.

- 1 Present ***Linkster***
A tool to facilitate link reverse-engineering.
- 2 Evaluate this tool
- 3 Analyze the comprehensive data set
Apache HTTP web server project



The Missing Links: Bugs and Bug-fix Commits

2010 FSE/ESEC

Linkster

Commit Information

Revision	Date	Committer	Message
291120	2005-09-23	jorton	* server/request.c (core_opts_merge): When AllowOverride is ...
291123	2005-09-23	bnicholes	mod_lldap: Fix PR 36563. Keep track of the number of attribut...
291125	2005-09-23	jorton	Merge r291120 from trunk:
291128	2005-09-23	jorton	Fix entry format and remove a docs change which is not really
291265	2005-09-24	nd	update transformation

Merge r291120 from trunk:

* server/request.c (core_opts_merge): When AllowOverride is specified for the directory, ignore the inherited override_opts field.

PR: 35330
Submitted by: kabe <kabe@sra-tohoku.co.jp>
Reviewed by: jorton

Filename
/httpd/httpd/branches/2.2.x/server/request.c
/httpd/httpd/branches/2.2.x/CHANGES

Annotations

- ☐ Maintenance
- ☐ Feature
- ☐ Other
- ☒ Bug Fix

Bug IDs
35330

Notes
Backport of bug fix from trunk into 2.2.x branch

a

Blame Diff for Commit 97380

commit 2ca32d8d38c7b2a03db15f1571ebb7fff1947330
Author: William A. Rowe Jr <wrowe@apache.org>
Date: Sat Oct 6 22:16:11 2001 +0000

Aaargh! Unwinding part of my patch before I committed the prior version, I ended up blasting these (intentional) changes as well :(

Before Commit	After Commit
91342 int sec_idx;	91342 int sec_idx;
91342 int matches = cache	91342 int matches = cache
91342 walk_walked_t *last	91342 walk_walked_t *last
91342 core_dir_config *th	91342 core_dir_config *th
- 91342 allow_options_t opt	+ 97380 core_opts_t opts;
- 91342 allow_options_t opt	
- 91342 allow_options_t opt	
- 91342 overrides_t override	
91816 apr_finfo_t thisinfo	91816 apr_finfo_t thisinfo
91816 char *save_path_info	91816 char *save_path_info
91342 apr_size_t buflen;	91342 apr_size_t buflen;
91342 char *buf;	91342 char *buf;
91342 unsigned int seq, s	91342 unsigned int seq, s

c

Bug Information

Bug ID	Date	Description
26467	2005-09-16	, then hangs long-lived httpd children
25659	2005-09-21	Memory leak in ssl_util_algotypeof().
10470	2005-09-23	proxy module will not correctly serve mixed case fi...
35330	2005-09-24	[PATCH] AllowOverride Options does not allow Opti...
36563	2005-09-24	mod_lldap caching breaks when value is NULL
35292	2005-09-26	ap_lingering_close does not linger up to MAX_SECS...
35438	2005-09-26	--enable-static-rotatlogs still dynamically links ...
12355	2005-09-28	POST incompatible w/ renegotiate https: connection
32037	2005-09-28	mod_ssl.h in err.h no
36424	2005-10-03	AutoIndex fa after reboot

Extended Bug Entry Bug Activity Fixing Files

Status: RESOLVED

Reported By: kabe <kabe@sra-tohoku.co.jp>
2005-06-12 08:16:56

Assigned To: Apache HTTPD Bugs M

Severity: regression

Priority: P2

OS: All

Product: Apache httpd-2

Component: Core

Platform: All

Version: 2.1.7

Resolution: FIXED

Descriptions

kabe@sra-tohoku.co.jp
2005-06-12 08:16:56

*** Bug 35329 has been marked as a duplicate of this bug. ***

sander@temme.net 2005-09-20 05:47:51

We now have a 'regression' severity in Bugzilla. Removed the keyword.

jorton@redhat.com 2005-09-24 00:29:46

Thanks a lot for the patch (attaching it might have avoided the line-wrapping ;). This has been applied for 2.1.8.

<http://svn.apache.org/viewcvsr>

b



The Missing Links: Bugs and Bug-fix Commits

2010 FSE/ESEC

- RQ 1: *Do the bug reporting and fixing practices of developers correspond to the assumptions commonly made by researchers?*

A so-called “bug” is not always a bug; neither is a “commit” always a commit.

- RQ 2: *How well does the automated approach of finding links between commits and bug reports work?*

The automated approach finds virtually all the commit log messages which contain a link to the bug tracking database

- RQ 3: *Is there any evidence of systematic bias in the linking of bug-fix commits to bug reports?*

Find that reporting bias affects the performance of a bug prediction algorithm .



The Missing Links: Bugs and Bug-fix Commits

2010 FSE/ESEC

Finding 1. Not all fixed bugs are mentioned in the bug tracking database. Some are discussed (only) on the mailing list.

Finding 2. To fix a bug in an Apache release, multiple similar commits by different developers are needed.

Finding 3. Developers sometimes fix bugs that are only reported in some other projects' bug tracker, rather than in their own; and vice-versa.

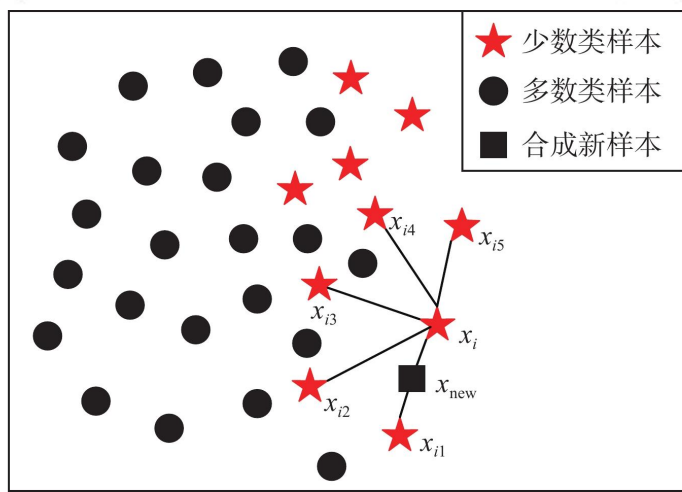
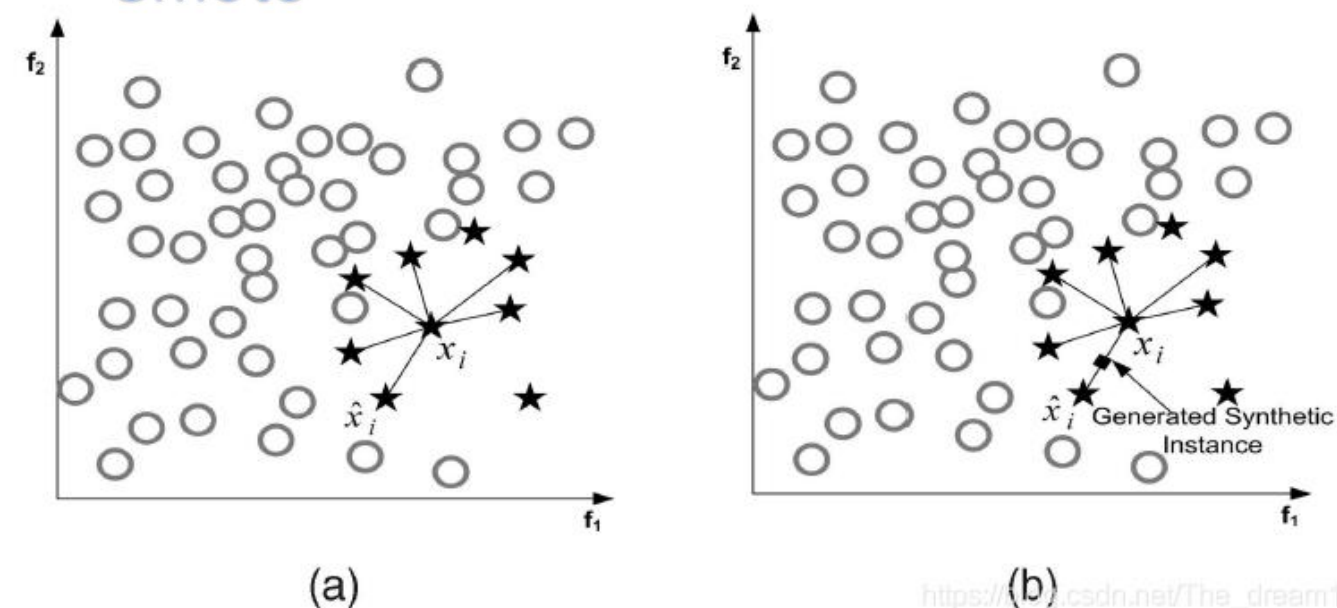
Finding 4. Even if we annotate all commits, the cause of a commit still remains unspecified in some cases.



Is better data better than better data miners?: on the benefits of tuning smote for defect prediction

2018 ICSE

Smote



```
def SMOTE(k=2, m=50%, r=2): # defaults
    while Majority > m do
        delete any majority item # random
    while Minority < m do
        add something_like(any minority item)
```

```
def something_like(X0):
    relevant = emptySet
    k1 = 0
    while(k1++ < 20 and size(found) < k) {
        all = k1 nearest neighbors
        relevant += items in "all" of X0 class}
    Z = any of found
    Y = interpolate (X0, Z)
    return Y
```

```
def minkowski_distance(a,b,r):
    return  $(\sum_i abs(a_i - b_i)^r)^{1/r}$ 
```

Figure 3: Pseudocode of SMOTE



Is better data better than better data miners?: on the benefits of tuning smote for defect prediction

2018 ICSE

Smotuned

```
def DE( n=10, cf=0.3, f=0.7): # default settings
    frontier = sets of guesses (n=10)
    best = frontier.1 # any value at all
    lives = 1
    while(lives-- > 0):
        tmp = empty
        for i = 1 to |frontier|: # size of frontier
            old = frontieri
            x,y,z = any three from frontier, picked at random
            new= copy(old)
            for j = 1 to |new|: # for all attributes
                if rand() < cf # at probability cf...
                    new.j = x.j + f * (z.j - y.j) # ...change item j
            # end for
            new = new if better(new,old) else old
            tmpi = new
            if better(new,best) then
                best = new
                lives++ # enable one more generation
            end
        # end for
        frontier = tmp
    # end while
    return best
```

Table 5: SMOTE parameters

Para	Defaults used by SMOTE	Tuning Range (Explored by (SMOTUNED)	Description
<i>k</i>	5	[1,20]	Number of neighbors
<i>m</i>	50%	{50, 100, 200, 400}	Number of synthetic examples to create. Expressed as a percent of final training data.
<i>r</i>	2	[0.1,5]	Power parameter for the Minkowski distance metric.

Figure 4: SMOTUNED uses DE (differential evolution).



Is better data better than better data miners?: on the benefits of tuning smote for defect prediction

2018 ICSE

- RQ1: Are the default “off-the-shelf” parameters for SMOTE appropriate for all datasets?

- RQ2: Is there any benefit in tuning the default parameters of SMOTE for each new dataset?

- RQ3: In terms of runtimes, is the cost of running SMOTUNED worth the performance improvement?

- RQ4: How does SMOTUNED perform against more recent class imbalance technique?



Is better data better than better data miners?: on the benefits of tuning smote for defect prediction

2018 ICSE

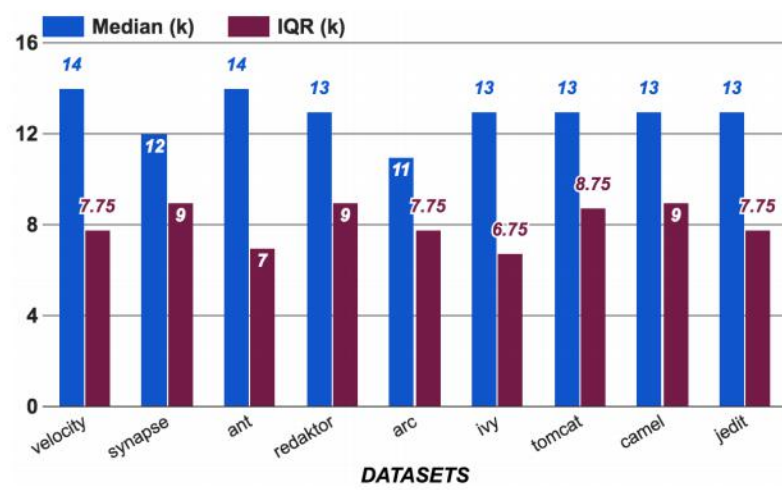


Figure 5a: Tuned values for k (default: $k = 5$).

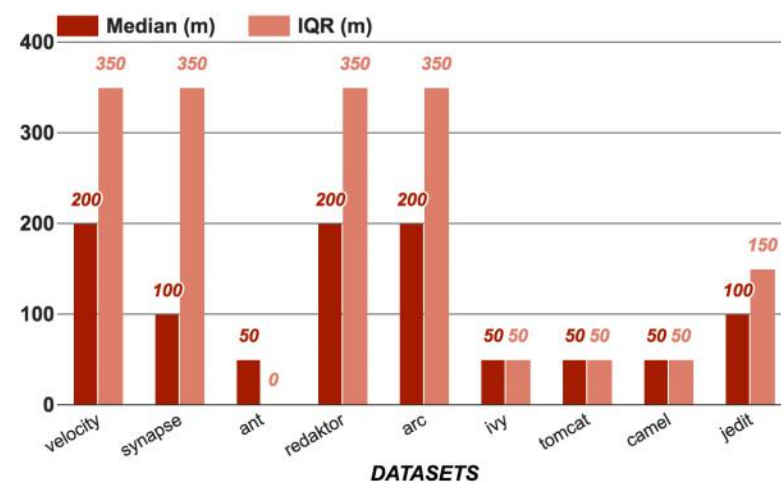


Figure 5b: Tuned values for m (default: $m = 50\%$).

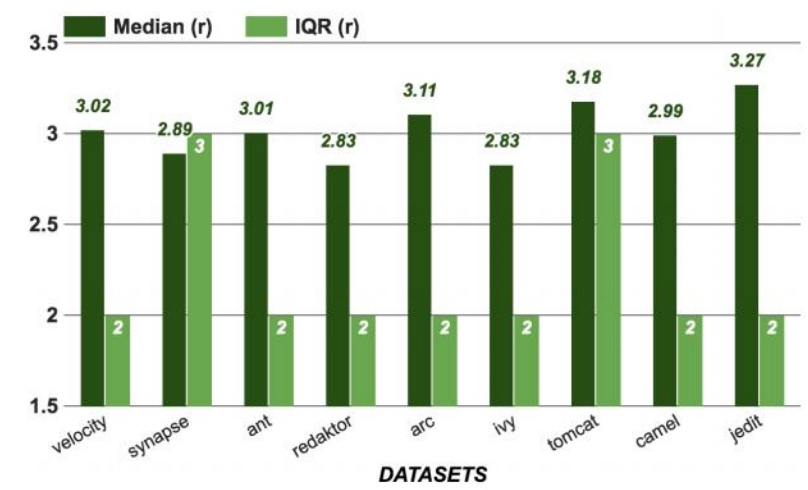


Figure 5c: Tuned values for r (default: $r = 2$).

Figure 5: Data sets vs Parameter Variation when optimized for recall and results reported on recall. “Median” denotes 50th percentile values seen in the 5*5 cross-validations and “IQR” shows the intra-quartile range, i.e., (75-25)th percentiles.

Treatments	AUC	number of wins		
		Recall	Precision	False Alarm
MAHAKIL	1/9	0/9	6/9	9/9
SMOTE	0/9	1/9	0/9	0/9
SMOTUNED	8/9	8/9	3/9	0/9



Is better data better than better data miners?: on the benefits of tuning smote for defect prediction

2018 ICSE

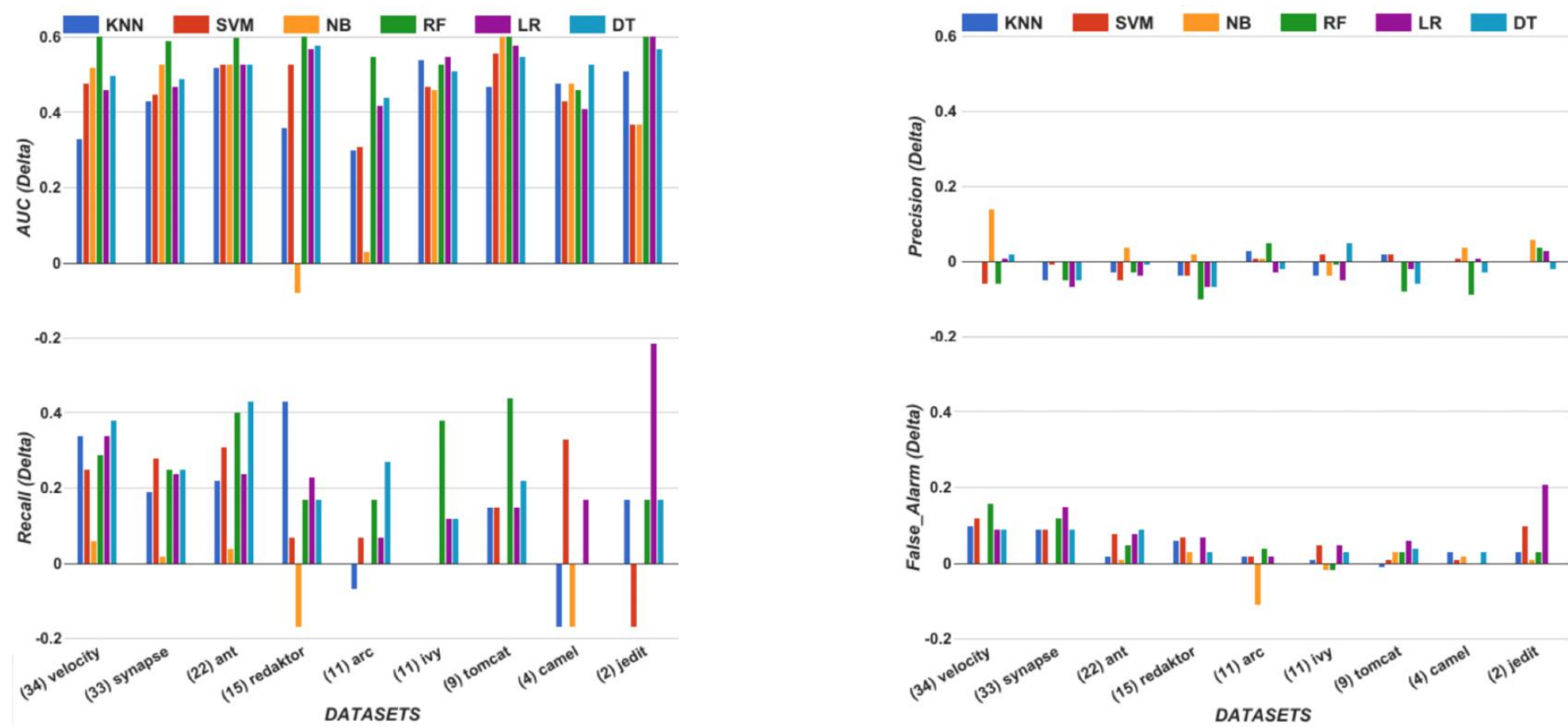


Figure 6: SMOTUNED improvements over SMOTE. Within-Measure assessment (i.e., for each of these charts, optimize for performance measure M_i , then test for performance measure M_i). For most charts, *larger* values are *better*, but for false alarm, *smaller* values are *better*. Note that the corresponding percentage of minority class (in this case, defective class) is written beside each data set.