SCEADAN v1.0

Systematic Classification Engine and Data ANalysis

Overview by Dr. Nicole Beebe

The University of Texas at San Antonio

June 30, 2012

The UTSA Team…

Laurence Maddox
CS Undergrad Student
(Software Development)

Lishu Liu
IT Grad Student
(SVM Experiments)

DJ Bauch
Chief Scientist, SAIC IO Div.
(S/W Dev. Mentorship)
Seen here mentoring San Antonio high school students in robotics competition

Nicole Beebe
Asst. Prof., UTSA
(Theory, Design,
Data Prep & Analysis)

Experimental Process

- Training data collection & preparation
  - Identify target data types
    - Challenge identified 38 types
    - Added 4 types prevalent to Govdocs dataset (.ps, ~~.pps~~, .bmp, .java)
  - Collect sample files

40% from Govdocs v1.1

.text	.png	.ppt	.bmp
.csv	.gif	.docx	.java
.log	.gz	.xlsx	
.html	.pdf	.pptx	
.xml	.doc	.ps	
.jpg	.xls	.pps	

60% researcher collected

.txt	.bz2	.flv	Hex enc.
.css	.xlsx	FAT	Encryp.
.js	.mp3	NTFS	Random
.json	.m4a	Ext3	constant
.tiff	.avi	Base64	
.zip	.wmv	Base85	

NOTE: Items in red indicate potential mischaracterization issues in Govdocs v1.1 dataset

Experimental Process (cont.)

- Data preparation
  - Verified data type via “signature analysis”
    - Looked for extension – signature match (discarded otherwise)
  - Segmented files (512B blocks)
  - Removed header segments
  - Resulted in 1,042,027 fragments for experimentation
- Selected classification mechanism … support vector machine
- Feature identification and extraction
  - Literature review to select features (next slides…)
  - Calculate normalized feature values
  - Create libsvm/liblinear formatted vectors
  - Cross-validation training for parameter selection

References

- Fitzgerald et al. (2012)
- Gopal et al. (2011)
- Axelsson (2010)
- Conti et al. (2010)
- Li et al. (2010)
- Ahmed et al. (2010)
- Ahmed et al. (2009)
- Calhoun and Coles (2008)
- Moody and Erbacher (2008)
- Veenman (2007)
- Erbacher and Mulholland (2007)
- Karresand and Shahmehri (2006)
- Hall and Davis (2006)
- Li et al. (2005)
- McDaniel and Heydari (2003)
- Shannon (2004)

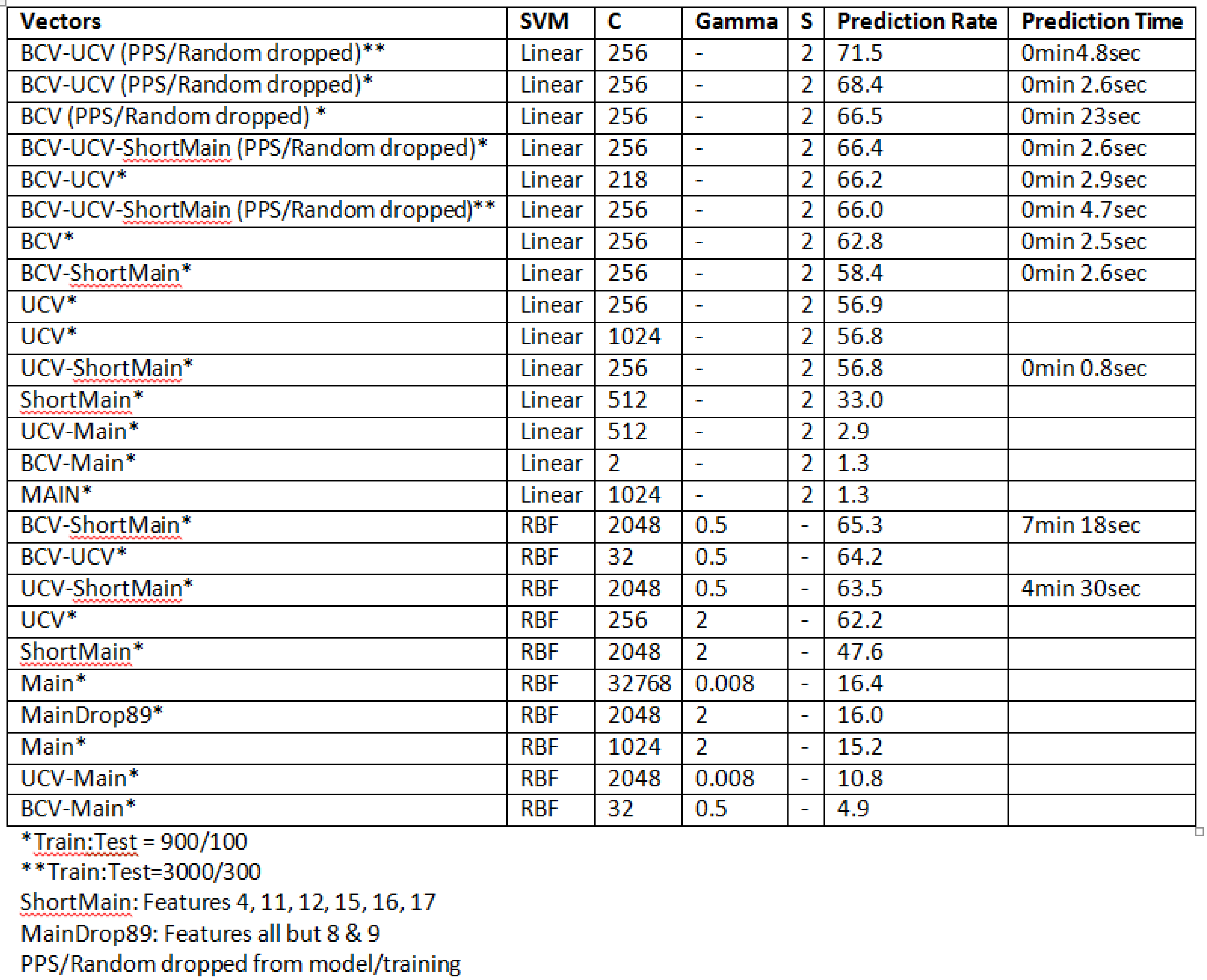
Features Identified (BCV+UCV = final model)

- **Unigram count vector (UCV)**
- **Bigram count vector (BCV)**
- Bi-gram entropy
- Item entropy
- Hamming weight
- Mean byte value
- Standard dev. of byte values
- Kurtosis
- Max byte streak
- Avg. contiguity between bytes
- Compressed item length
  - Burrows-Wheeler
  - LZW
- ASCII frequencies
  - Low (0x00-0x1F)
  - Med (0x20-0x7F)
  - High (0x80-0xFF)
- Byte value correlation
- Byte value frequency correlation

grey= not coded in; others are coded, but disabled in v1.0 because not used in model

“UCV+BCV” here means vector concatenation

Experimental Results Summary (25 exp. variations)



Vectors	SVM	C	Gamma	S	Prediction Rate	Prediction Time
BCV-UCV (PPS/Random dropped)**	Linear	256	-	2	71.5	0min4.8sec
BCV-UCV (PPS/Random dropped)*	Linear	256	-	2	68.4	0min 2.6sec
BCV (PPS/Random dropped) *	Linear	256	-	2	66.5	0min 23sec
BCV-UCV-ShortMain (PPS/Random dropped)*	Linear	256	-	2	66.4	0min 2.6sec
BCV-UCV*	Linear	218	-	2	66.2	0min 2.9sec
BCV-UCV-ShortMain (PPS/Random dropped)**	Linear	256	-	2	66.0	0min 4.7sec
BCV*	Linear	256	-	2	62.8	0min 2.5sec
BCV-ShortMain*	Linear	256	-	2	58.4	0min 2.6sec
UCV*	Linear	256	-	2	56.9	
UCV*	Linear	1024	-	2	56.8	
UCV-ShortMain*	Linear	256	-	2	56.8	0min 0.8sec
ShortMain*	Linear	512	-	2	33.0	
UCV-Main*	Linear	512	-	2	2.9	
BCV-Main*	Linear	2	-	2	1.3	
MAIN*	Linear	1024	-	2	1.3	
BCV-ShortMain*	RBF	2048	0.5	-	65.3	7min 18sec
BCV-UCV*	RBF	32	0.5	-	64.2	
UCV-ShortMain*	RBF	2048	0.5	-	63.5	4min 30sec
UCV*	RBF	256	2	-	62.2	
ShortMain*	RBF	2048	2	-	47.6	
Main*	RBF	32768	0.008	-	16.4	
MainDrop89*	RBF	2048	2	-	16.0	
Main*	RBF	1024	2	-	15.2	
UCV-Main*	RBF	2048	0.008	-	10.8	
BCV-Main*	RBF	32	0.5	-	4.9	

*Train:Test = 900/100

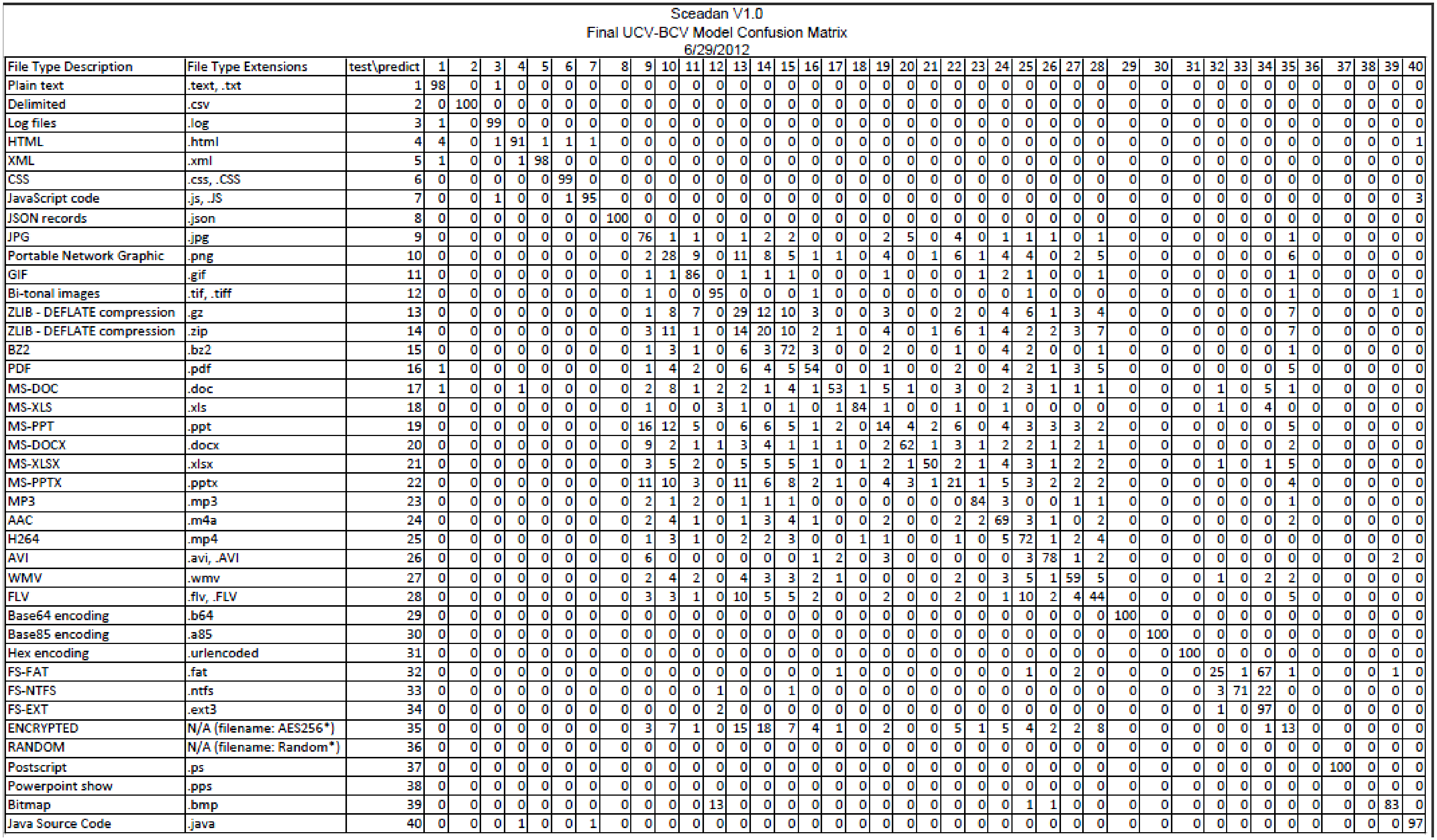
**Train:Test=3000/300

ShortMain: Features 4, 11, 12, 15, 16, 17

MainDrop89: Features all but 8 & 9

PPS/Random dropped from model/training

Final Confusion Matrix



Sceadan V1.0
Final UCV-BCV Model Confusion Matrix
6/29/2012

File Type Description	File Type Extensions	test\predict	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Plain text	.text, .txt	1	98	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Delimited	.csv	2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Log files	.log	3	1	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HTML	.html	4	4	0	1	91	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
XML	.xml	5	1	0	0	1	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CSS	.css, .CSS	6	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JavaScript code	.js, .JS	7	0	0	1	0	0	1	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
JSON records	.json	8	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JPG	.jpg	9	0	0	0	0	0	0	0	0	76	1	1	0	1	2	2	0	0	0	2	5	0	4	0	1	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0
Portable Network Graphic	.png	10	0	0	0	0	0	0	0	0	2	28	9	0	11	8	5	1	1	0	4	0	1	6	1	4	4	0	2	5	0	0	0	0	0	0	6	0	0	0	0	0
GIF	.gif	11	0	0	0	0	0	0	0	0	1	1	86	0	1	1	1	0	0	0	1	0	0	0	1	2	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
Bi-tonal images	.tif, .tiff	12	0	0	0	0	0	0	0	0	1	0	0	95	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
ZLIB - DEFLATE compression	.gz	13	0	0	0	0	0	0	0	0	1	8	7	0	29	12	10	3	0	0	3	0	0	2	0	4	6	1	3	4	0	0	0	0	0	0	7	0	0	0	0	0
ZLIB - DEFLATE compression	.zip	14	0	0	0	0	0	0	0	0	3	11	1	0	14	20	10	2	1	0	4	0	1	6	1	4	2	2	3	7	0	0	0	0	0	0	7	0	0	0	0	0
BZ2	.bz2	15	0	0	0	0	0	0	0	0	1	3	1	0	6	3	72	3	0	0	2	0	0	1	0	4	2	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
PDF	.pdf	16	1	0	0	0	0	0	0	0	1	4	2	0	6	4	5	54	0	0	1	0	0	2	0	4	2	1	3	5	0	0	0	0	0	0	5	0	0	0	0	0
MS-DOC	.doc	17	1	0	0	1	0	0	0	0	2	8	1	2	2	1	4	1	53	1	5	1	0	3	0	2	3	1	1	1	0	0	0	1	0	5	1	0	0	0	0	0
MS-XLS	.xls	18	0	0	0	0	0	0	0	0	1	0	0	3	1	0	1	0	1	84	1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	4	0	0	0	0	0	0
MS-PPT	.ppt	19	0	0	0	0	0	0	0	0	16	12	5	0	6	6	5	1	2	0	14	4	2	6	0	4	3	3	3	2	0	0	0	0	0	0	5	0	0	0	0	0
MS-DOCX	.docx	20	0	0	0	0	0	0	0	0	9	2	1	1	3	4	1	1	1	0	2	62	1	3	1	2	2	1	2	1	0	0	0	0	0	0	2	0	0	0	0	0
MS-XLSX	.xlsx	21	0	0	0	0	0	0	0	0	3	5	2	0	5	5	5	1	0	1	2	1	50	2	1	4	3	1	2	2	0	0	0	1	0	1	5	0	0	0	0	0
MS-PPTX	.pptx	22	0	0	0	0	0	0	0	0	11	10	3	0	11	6	8	2	1	0	4	3	1	21	1	5	3	2	2	2	0	0	0	0	0	0	4	0	0	0	0	0
MP3	.mp3	23	0	0	0	0	0	0	0	0	2	1	2	0	1	1	1	0	0	0	0	0	0	0	84	3	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0
AAC	.m4a	24	0	0	0	0	0	0	0	0	2	4	1	0	1	3	4	1	0	0	2	0	0	2	2	69	3	1	0	2	0	0	0	0	0	0	2	0	0	0	0	0
H264	.mp4	25	0	0	0	0	0	0	0	0	1	3	1	0	2	2	3	0	0	1	1	0	0	1	0	5	72	1	2	4	0	0	0	0	0	0	0	0	0	0	0	0
AVI	.avi, .AVI	26	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	1	2	0	3	0	0	0	0	0	3	78	1	2	0	0	0	0	0	0	0	0	0	0	2	0
WMV	.wmv	27	0	0	0	0	0	0	0	0	2	4	2	0	4	3	3	2	1	0	0	0	0	2	0	3	5	1	59	5	0	0	0	1	0	2	2	0	0	0	0	0
FLV	.flv, .FLV	28	0	0	0	0	0	0	0	0	3	3	1	0	10	5	5	2	0	0	2	0	0	2	0	1	10	2	4	44	0	0	0	0	0	0	5	0	0	0	0	0
Base64 encoding	.b64	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
Base85 encoding	.a85	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
Hex encoding	.urlencoded	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
FS-FAT	.fat	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	2	0	0	0	0	25	1	67	1	0	0	0	1	0
FS-NTFS	.ntfs	33	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	71	22	0	0	0	0	0	0
FS-EXT	.ext3	34	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	97	0	0	0	0	0	0
ENCRYPTED	N/A (filename: AES256*)	35	0	0	0	0	0	0	0	0	3	7	1	0	15	18	7	4	1	0	2	0	0	5	1	5	4	2	2	8	0	0	0	0	0	1	13	0	0	0	0	0
RANDOM	N/A (filename: Random*)	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Postscript	.ps	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
Powerpoint show	.pps	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bitmap	.bmp	39	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	83	0
Java Source Code	.java	40	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97

NOTE: Random & PPS removed from svm model; thus row of all zeroes

True Positive Prediction Rates in our Experiments

Type	Ext	Rate %
Delimited	.csv	100
JSON records	.json	100
Base64 encoding	.b64	100
Base85 encoding	.a85	100
Hex encoding	.urlenc	100
Postscript	.ps	100
Log files	.log	99
CSS	.css	99
Plain text	.text, .txt	98
XML	.xml	98
FS-EXT	.ext3	97
Java Source Code	.java	97
JavaScript code	.js	95
Bi-tonal images	.tif, .tiff	95
HTML	.html	91
GIF	.gif	86
MS-XLS	.xls	84
MP3	.mp3	84
Bitmap	.bmp	83
AVI	.avi	78
JPG	.jpg	76
BZ2	.bz2	72
H264	.mp4	72
FS-NTFS	.ntfs	71

Type	Ext	Rate %
AAC	.m4a	69
MS-DOCX	.docx	62
WMV	.wmv	59
PDF	.pdf	54
MS-DOC	.doc	53
MS-XLSX	.xlsx	50
FLV	.flv, .FLV	44
ZLIB – DEFLATE	.gz	29
Portable Network Graphic	.png	28
FS-FAT	.fat	25
MS-PPTX	.pptx	21
ZLIB - DEFLATE	.zip	20
MS-PPT	.ppt	14
ENCRYPTED	N/A	13

Average Sceadan prediction accuracy:
71.5%

Random chance classification:
1/40 = 2.5%

Train/Test: 3,000/300

Challenge Report Results – Initial Reflection

Sceadan v1.0 had 36% overall prediction accuracy on Challenge test data

Expectedly Poor Classifiers

- Need better training data
  - Audio (mp3, mp4)
  - Video (wmv, h264)
- Trouble with high entropy
  - Images (jpg, gif, png)
  - Office 2010 (docx, xlsx*, pptx*)
  - Lossless comp. (zlib, bzip)
- PDF - High encoding variety

Surprisingly Poor Classifiers

- Base16 encoding
  - We trained on URL encoding, not pure hex encoding!
- Markup (xml, html)
  - We did not filter out scripts
- Delimited text files
  - Not sure re: csv files!
  - Didn’t train on other separators

* Did surprisingly *well* on .xlsx and .pptx

Miscellaneous

- To get/run … download and make
  - http://www.sceadan.com/code/sceadan.v1.tar.bzip2
    - Need liblinear v1.91 (http://www.csie.ntu.edu.tw/~cjlin/liblinear)
  - Readme for usage instructions
- Tested on a few linux distros
- Language: C
- Copyright: University of Texas at San Antonio
- License: GPLv2

More Miscellaneous

- Other capabilities
  - Non-prediction mode
    - Generates libsvm compliant doc/block vectors
  - Can use your own libsvm model file
    - BCV+UCV based unless change code
  - Directory mode
- Fast, but not yet threaded
- Why “Sceadan”??
  - Old English / Proto-Germanic for “to classify”

Acknowledgements

- Student Funding
  - NPS Grant No. N00244-11-1-0011, “Advanced Digital Forensic String Search Capability”
    - Needed data type classifier for unallocated blocks for search hit ranking algorithm
  - UTSA Provost’s Summer Research Mentorship Program
    - Selects ~10 undergrad/professor pairs for summer research
    - Student funded 30 hrs/week, GRE prep/test, other training
    - Goal is to prepare select undergrads for grad or Ph.D. studies

Nicole.Beebe@utsa.edu

COMMENTS / QUESTIONS ?

Information about Researcher Created Training Data

BACK-UP SLIDES

Researcher Created Training Data

- TXT: Project Gutenberg ebooks
- BZ2: Ubuntu’s bzip2 utility on .txt files
- Base64: Ubuntu’s uuencode utility on Govdocs .zip/.gz files
- Base85: Online encoder (webutils.pl) of bzipped .txt files
- Hex Encoding: Online encoder (webutils.pl) of .txt files
- CSS: css templates from web, personal PC files
- JS: sourceforge.net
- JSON: Personal PC files, CSV files converted to JSON via web utility (johntron.com)
- TIF/TIFF: Adobe Pro9 save personal files as CCITT-G3/G4
- ZIP: Windows7 zip utility on .txt files

Researcher Created Training Data (cont.)

- XLSX: Added some personally created files
- MP3: 10 personally owned audio files
- AAC: 10 personally owned audio files
- WMV: 28 files from Windows install
- AVI: 13 files from Windows install
- FLV: Personally owned files and downloaded from www
- MP4: Personally owned files and downloaded from www
- Encrypted: AES256 bitlockered data from personal PC
- Random: random.org

Researcher Created Training Data (cont.)

- FAT
  - Types: FATs, Dir Entry Structures, VBR
  - Sources: USBNIST1, Gen3, FAT32 from digitalcorpora.org, plus one researcher created FAT32 image
- NTFS
  - Types: $AttrDef, $BadClus, $Bitmap, $Boot, $LogFile, $MFT, $MFTMirr, $Secure, $UpCase, $Volume
  - Source: “Patents” images (last day) digitalcorpora.org
- EXT
  - Types: Block descriptor, extent block, group descriptor, inode bitmap, inode table, journal area, volume bitmap
  - Source: Casper-RW, Gen3, EXT3 images from digitalcorpora.org