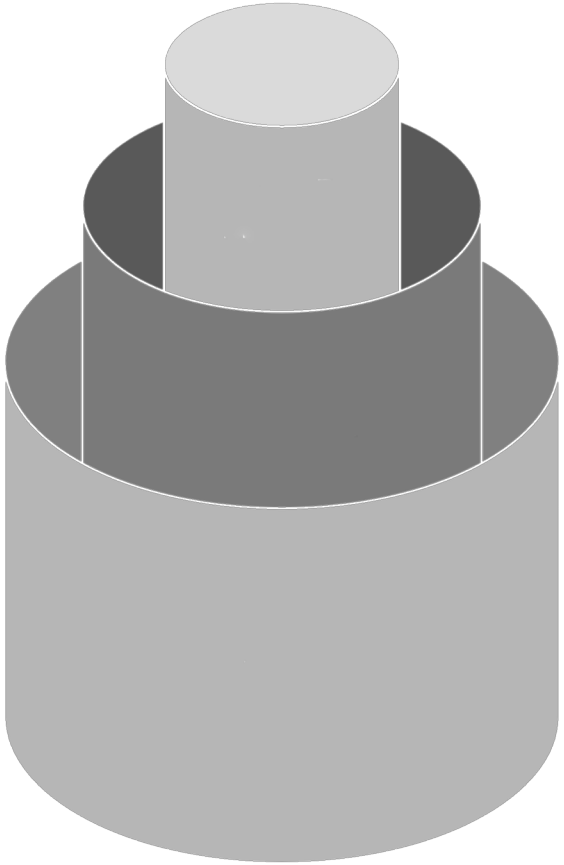


Makine Öğrenmesi için Büyük Veri Kümelerinin Kesitinin Çıkartılmasında Yeni Teknikler:

Android Mobil Kötücül Veri Kümelerinin Özlü Bir Gözden Geçirilmesi*

Gürol Canbek
Orta Doğu Teknik Üniversitesi
http://gurol.canbek.com

Şeref Sağıroğlu
Gazi Üniversitesi

Tuğba Taşkaya Temizel
Orta Doğu Teknik Üniviersitesi

*Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, “New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets”, International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

Özet

SORUN:

- Büyük veri boyutları:
  - Çokluk, Çabukluk, Çeşitlilik
  - Dürüstlük, Değer, Değişkenlik
- Makine öğrenmesinde büyük verinin kullanımı?

KATKI:

- Bu çalışma veri kümelerinin kesitlerinin çıkartılması için
  - 14 kıstas ile 4 farklı teknik önermektedir.
  - İlk defa: Büyük veri boyutları ile eşleme

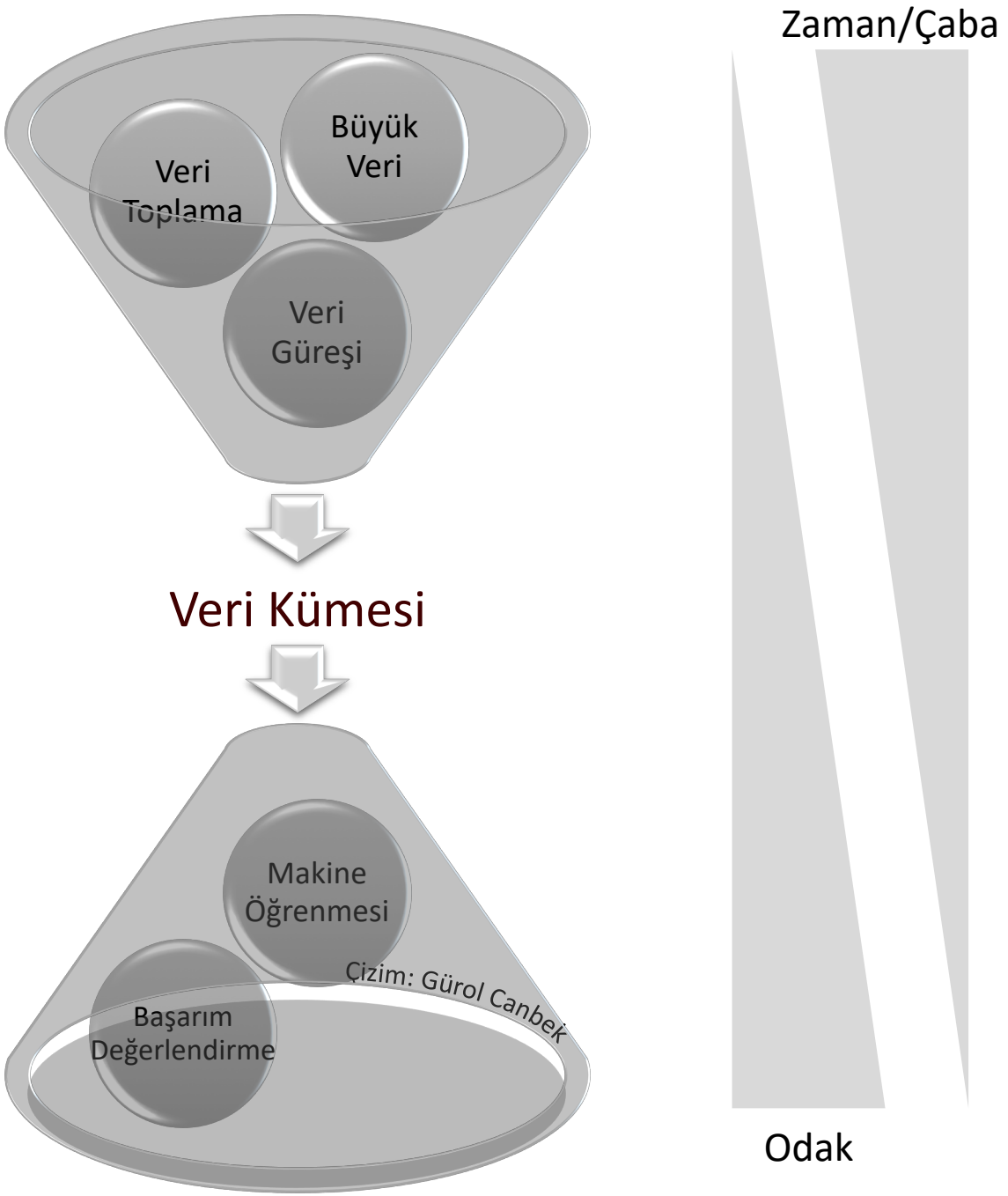
DENEY:

- Android Malware Genome Project, Drebin, Android Malware Dataset, Anroid Botnet ve Virus Total 2018

SONUÇ:

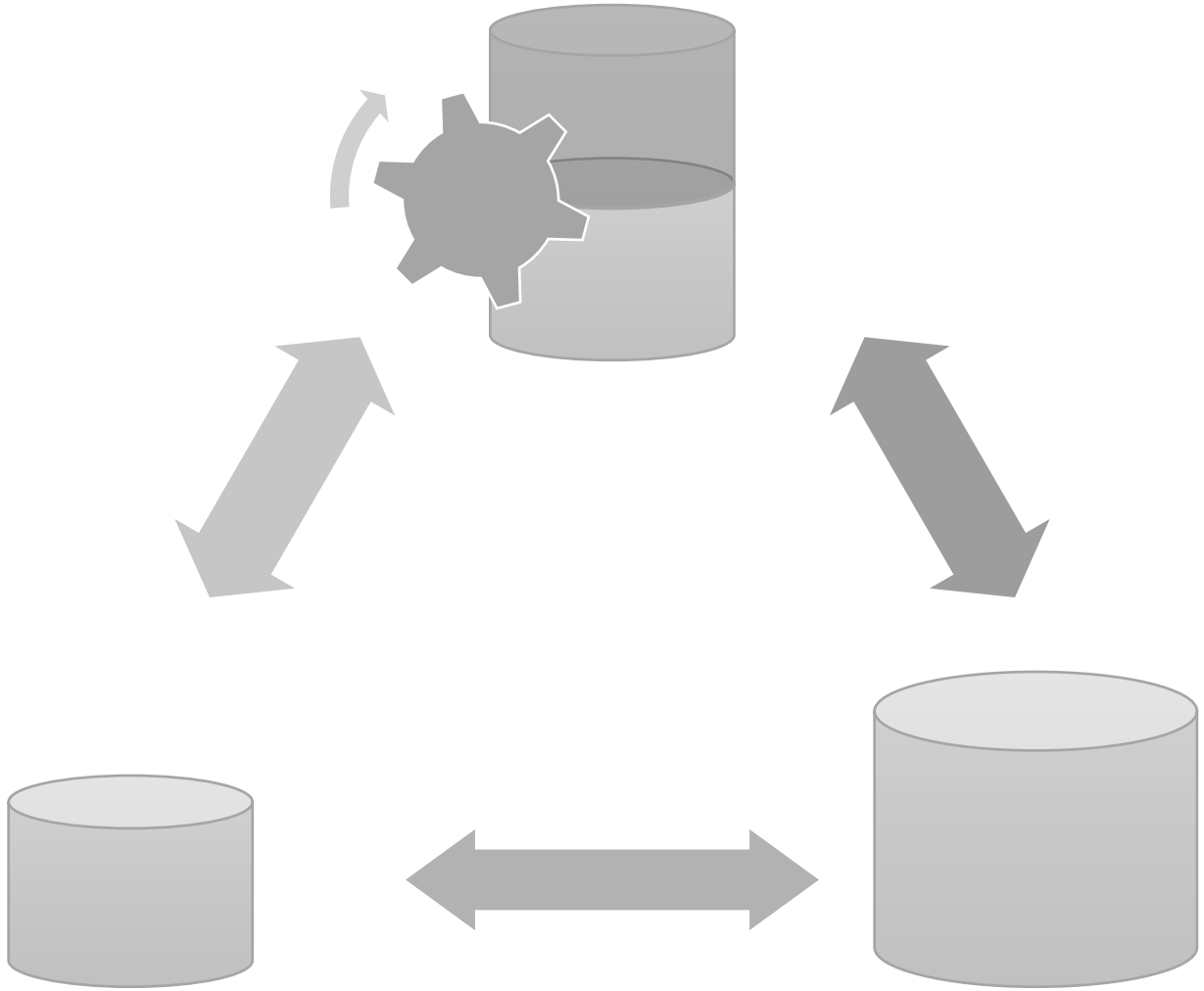
- Yöntem, veri kümeleri hakkında karşılaştırılabilir önemli içgörü sunmaktadır.
- Büyük verinin daha görünür, kaliteli ve özümsenmiş olmasına uygulanabilir bir katkı sağlamaktadır.

Büyük Veri ve Makine Öğrenme İş Akışı



Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, “New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets”, International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

Veri Kümelerinin Kalitesi



- Eldeki veri kümesi:
  - Kalite artırma
- Farklı veri kümeleri
  - Kalite karşılaştırma

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, “New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets”, International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

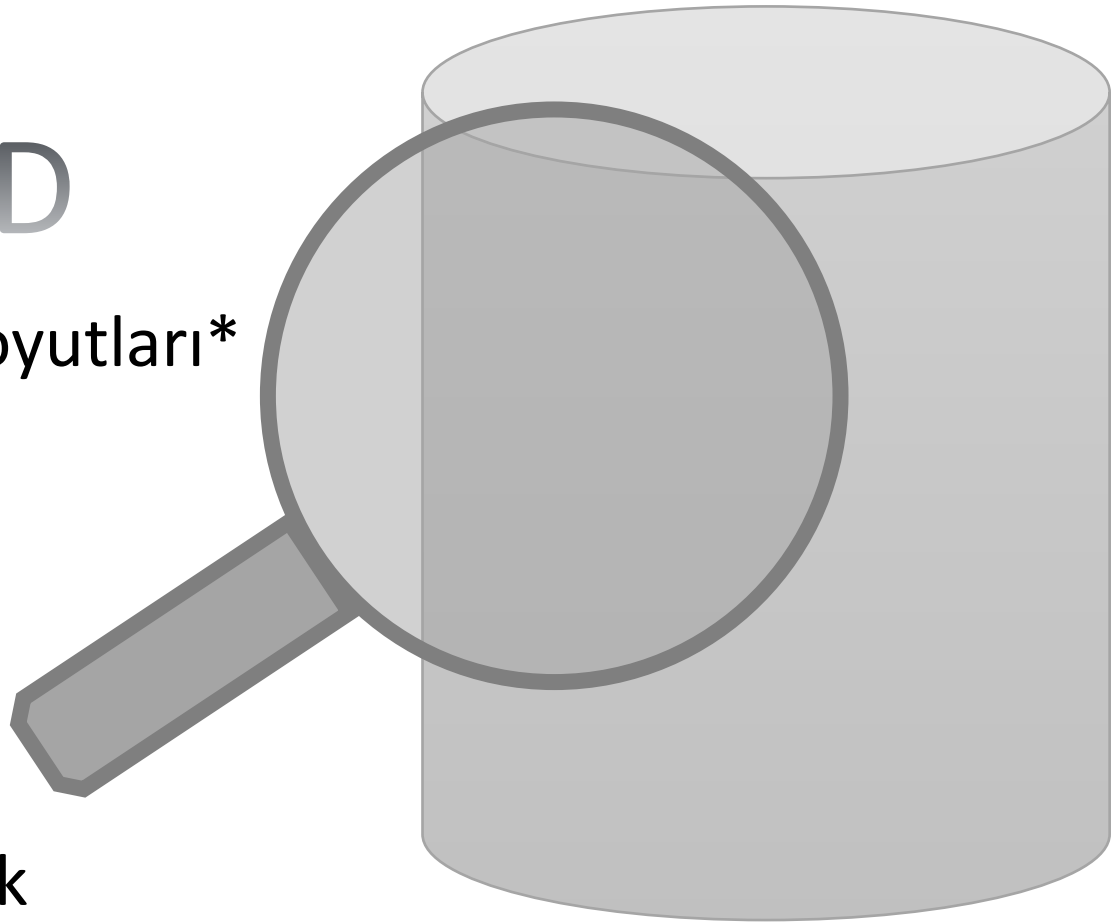
Veri Kümesi Kesit Çıkarma

- İlk planda
- Sistematik bir şekilde
- Önemli içgörü

3Ç, 4D

Büyük Veri Boyutları*

- Çokluk
- Çabukluk
- Çeşitlilik
- Dürüstlük
- Değer
- Değişkenlik
- Dayanak

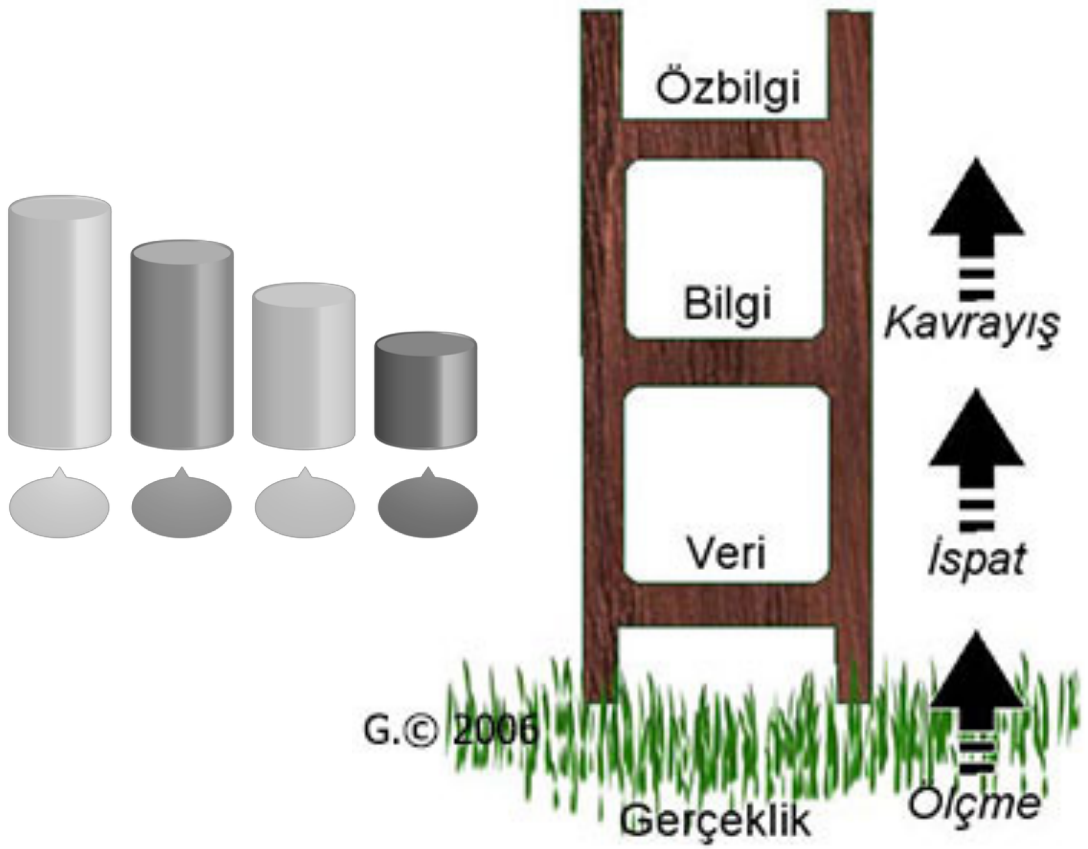


* Volume, Velocity, Variety, Veracity, Value, Variability, Venue için Türkçe karşılık önerileri: Gürol Canbek, Kasım 2018

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, “New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets”, International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

Literatür ve Veri/Bilginin Kalitesi

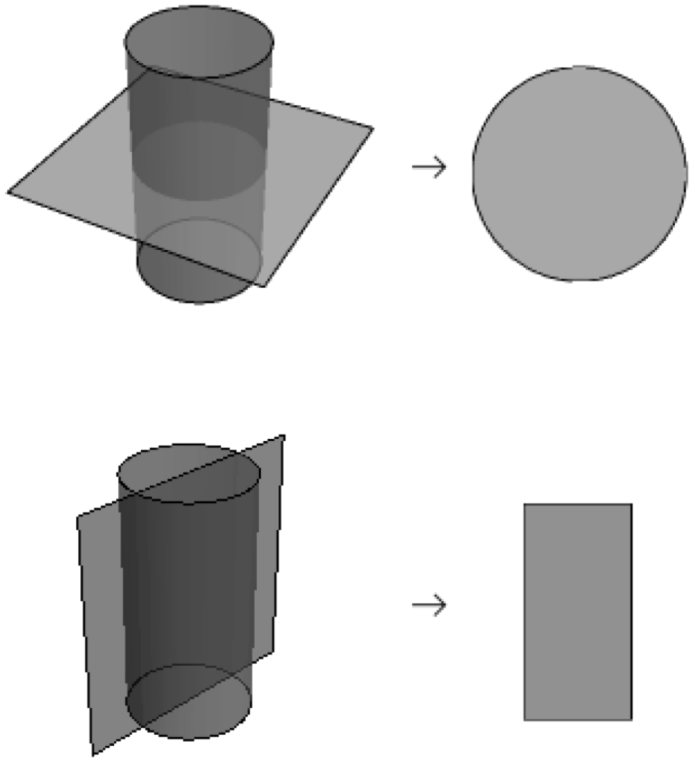
- Literatürde daha önce ele alınmış.
  - İçsel: Doğruluk, tutarlılık, …
  - Bağlamsal: İlgi, tamlık, …
  - Gösterimsel: Anlaşılırlık, biçem, …
  - Erişimsel: Kullanabilirlik, güvenlik, …
- İstatistiksel yöntemler
  - Tanımlayıcı istatistikler
- Sahaya özel kalite bakış açısı
  - Uzamsal veri kümeleri
  - Sağlık verileri
  - Doğa olayları verileri gibi



Canbek, G., Sağıroğlu, Ş. (2006). **Bilgi, Bilgi Güvenliği ve Süreçleri Üzerine Bir İnceleme**. *Politeknik Dergisi*, 9(3), 165–174.

Kesit Çıkarma

- Verinin veya veri kümesinin belirli bir bakış açısından tarif edilmesi.
- Makalede:
  - İlişkisel Veritabanı Sistemleri (RDBMS)
  - Yapısal Sorgulama Dilleri (SQL)
  - İnternet üzerindeki veriler
- Veri kümelerinin makine öğrenmesi açısından kalitesi?



Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, “New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets”, International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

Android Mobil Kötücül Yazılım Veri Kümeleri

Veri Kümesi (Kısaltma)	Yıl	Etüt*
Android Malware Genome Project (AMGP)	2013	%65
Drebin	2014	%22
Android Botnet (ABot)	2015	
Android Malware Dataset (AMD)	2017	
VirusTotal Academic Malware Samples (VT2018)	2018	

* 2009-2018 atmış çalışma. İlave malzeme: github.com/gurol/dsprofiling

Kesit Çıkarma Zemini

- MalWareHouse
  - Android kötücül yazılım çözümleme zemini
  - Farklı veri kümesi havuzlarının veri ambarı yaklaşımı ile birleştirilmesi
  - Kesit çıkarma bilgilerinin elde edilmesi ve görselleştirilmesi
  - Python, MongoDB, R

MalWareHouse

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, “New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets”, International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

Önerilen Kesit Çıkarma Zümreleri

1. Temel
2. Zaman çizgisi
3. Mükerrer örneklem
4. Yoğunluk/seyreklik

Temel Kesit Çıkarım

- Üst seviyeden ilk içgörü

Kıstas	Büyük Veri Boyutu	AMGP	Drebin	AMD	ABot	VT2018
Örneklem uzayı büyüklüğü (m)	Çokluk	1,260	5,555	23,743	1,929	4,725
Öznitelik uzayı büyüklüğü (n)		65	94	105	78	111
Fizikî büyüklük (GB)		1.5	6.8	58.1	2.6	18.4
Kötücül ailesi*	Çeşitlilik	49	> 20	71	14	Yok
Kötücül başka biçimleri*	Dürüstlük Dayanak	Yok	Yok	135	Yok	Yok

* Sahaya bağlı kıstaslar

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, "New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets", International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

Zaman Çizgisi Kesit Çıkarımı

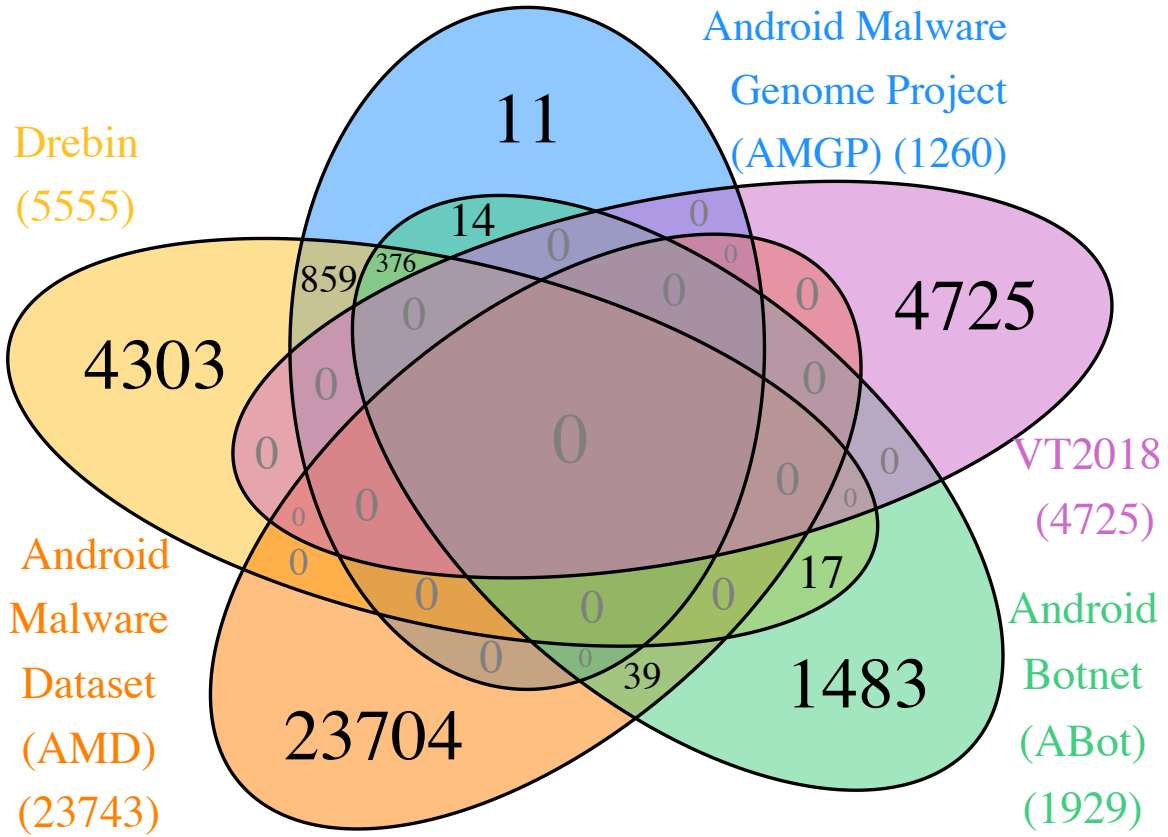
- **Yaş**: En genç ile en yaşlı örneklem arasındaki fark
- **Tazelik**: En genç örneklemden itibaren geçen zaman aşımı
- API seviyesi menzili (sahaya bağlı)
- Çabukluk, Dürüstlük, Değişkenlik
- Önerilen gösterim yöntemi:

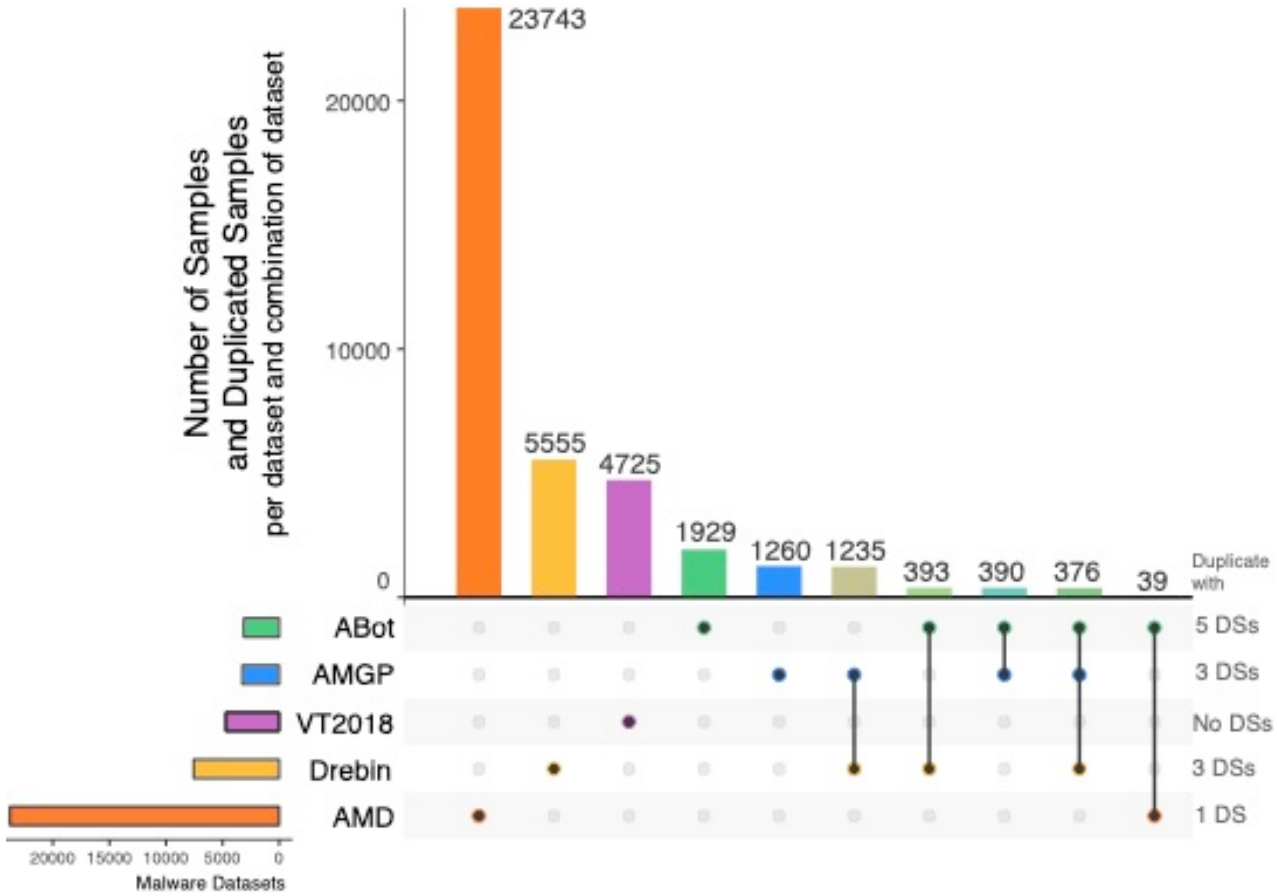
Veri Kümesi	En yaşlı	En genç	API	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Tazelik
AMGP	2008	2011	12	2.9 yaşında											-7 yıl
Drebin	2008	2012	16	4											-6
AMD	2008	2016	25	8											-2
ABot	2009	2015	21		7										-3
VT2018	2009	2018	27		9										0

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, "New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets", International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

Mükerrer Örneklem Kesit Çıkarımı

- Değer, Dayanak





Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, "New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets", International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

Yoğunluk/Seyreklik Kesit Çıkarımı

$$yoğunluk = \frac{gözlemlenen\ öznitelik\ sayısı}{m \times n}$$

$$seyreklik = \frac{gözlemlenmeyen\ öznitelik\ sayısı}{m \times n} = 1 - yoğunluk$$

$$yoğunluk_n = \frac{n}{n_{max}}$$

$$seyreklik_n = 1 - density_n$$

Kıstas	AMGP	Drebin	AMD	ABot	VT2018
Yoğunluk	18%	10%	11%	16%	15%
Seyreklik	82%	90%	89%	84%	85%
Öznitelik uzayı yoğunluğu	43%	62%	70%	52%	74%
Öznitelik uzayı seyrekliği	57%	38%	30%	48%	26%
Örneklem uzayı büyüklüğü (m)	1,260	5,555	23,743	1,929	4,725
Öznitelik uzayı büyüklüğü (n)	65	94	105	78	111

Çabukluk, Dürüstlük, Değer

Ayrıca, öznitelik uzayı yoğunluk ve seyrekliği (evrensel veya en büyük öznitelik uzayında göre)

Alınan Sonuçlar

- 14 kıstasın toplu değerlendirilmesi
- Yüksek kesit veri kümeleri:
  - Android Malware Dataset (AMD)
  - VirusTotal Academic Malware Samples (VT2018)
- Düşük kesit veri kümesi
  - Android Malware Genome Project (AMGP)

Kesit	Büyük Veri Boyutu[1]	Kıstas	AMGP	Drebin	AMD	ABot	VT2018
Temel	Çokluk	Örneklem uzayı büyüklüğü (m)	1260	5555	23743	1929	4725
		Öznitelik uzayı büyüklüğü (n)[2]	65	94	105	78	111
		Fizikî büyüklük (GB)	1.5	6.8	58.1	2.6	18.4
	Çeşitlilik, Dürüstlük, Dayanak	Kötücül ailesi[3]	49	> 20	71	14	N/A
		Kötücül başka biçimleri[3]	N/A	N/A	135	N/A	N/A
Zaman Çizgisi	Dürüstlük, Değişkenlik, Çabukluk	Yaş (yıl)	2.9	4.1	7.6	6.6	9
		Tazelik (yıl)	-7	-6	-2	-3	0
		API-seviyesi menzili[3]	12	16	25	21	27
Mükerrer Örneklem	Değer, Dayanak	Özgün örneklem sayısı	11	4303	23704	1483	4725
		Mükerrer örneklem sayısı	1249	1252	39	446	0
Yoğunluk / Seyreklik	Çokluk, Dürüstlük, Değer	Yoğunluk[4]	18%	10%	11%	16%	15%
		Seyreklik[4]	82%	90%	89%	84%	85%
		Öznitelik uzayı yoğunluğu	43%	62%	70%	52%	74%
		Öznitelik uzayı seyrekliği	57%	38%	30%	48%	26%
Genel sezgisel veri kümesi kesiti			Düşük (-5)	Normal (1)	Yüksek (4)	Normal (1)	Yüksek (4)

(1) Yedi boyut kapsanmıştır: İngilizce: value, variability, variety, velocity, venue, veracity, volume
(2) Öznitelik uzayı yoğunluğu/seyrekliği kıstasında değerlendirilmeye alınmıştır.
(3) Sahaya özel kıstas ancak diğer sahalara uygun bir şekilde uyarlanabilir
(4) Bilgi verici, herhangi bir üstünlük arz etmemektedir

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, "New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets", International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE

Değerlendirme ve Sonuçlar

Bu çalışmada;

- Dört zümrede ve 14 kıstastan oluşan veri kümesi kesit çıkartımı yöntemi önerilmiştir.
- İlk defa büyük veri boyutları veri kümesi kalitesi anlamında kıstaslarla eşleştirilmiştir.
- Kesitlerin anlaşılırlığını artırmak adına farklı görselleştirme yaklaşımları tanıtılmıştır.
- Yöntem, beş Android kötücü yazılım veri kümesinde denenmiş ve veri küme kesitlerinin karşılaştırılması yapılmıştır.

SONUÇ

- Veri kümelerinin ilk planda yüksek/düşük kesit şeklinde etiketlenmesi ilk planda elverişli bir içgörü sağlayabilir.
- Çalışmalarda kullanılan veri kümelerinin kesiti, başarım raporlamada ifade edilebilir.
- Araştırmacıları kaliteli veri kümesi elde etmek için teşvik sağlar.
- Yeni çıkan veri kümesi o saha içindeki veri kümelerinin yeniden kesitinin çıkartılmasını gerektirir.
- Bir veri kümesinin kesitinin çıkartılması eksik yönlerinin giderilmesini sağlar.

Teşekkürler

Gürol Canbek

http://gurol.canbek.com/Publications

Bildiri ile ilgili ilave bilgi, betik ve malzeme

https://github.com/gurol/dsprofiling

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, “New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets”, International Congress on Big Data, Deep Learning & Fighting Cyber Terrorism (IBIGDELFT 2018), 3–4 December 2018: IEEE