International Journal Of Computer Architecture And Mobility
(ISSN 2319-9229) Volume 1-Issue 2, December 2012

Research dataset Evaluation for Technologies

Shweta kori,Satyawan Mehra,Vishal Trivedi ,Romil Rawat

Shwetakori.250288@gmail.com,satyawan.mehra@gmail.com,vishalrtrivedi@gmail.com,rawat.romil@gmail.com

Abstract: Humans are free to think and create anything and that is correct from his point of view, but could not be applied globally or applied in existing systems. Research is that ,which are tested on global dataset and provides universal output and performance base and for future research also, if new research wants to modify it, As by own domains like Dshield is used when suspicious IP recommendation is required .Here in the propose work different dataset has been presented for strong analysis and research domains.

Introduction: Research comprises different parameters or designing descriptions like type of input and its features, output type and its feature, system[4] at which propose technique is evaluated by different parameters [1] like Accuracy percentage of system ,positive behavior like TPR(True Positive Rate) and FPR(False Positive rate), and Negative behavior like TNR(True Negative rate) and FNR(False Negative Rate).Training time defines how much time system takes to learns for taking decision, Detection time , defines time taken by system to detect behaviour,pattern,fingerprints,signatures etc.

Different type of input or dataset [2]is required by system to analyze and provide behavioral study. A dataset is a collection of different values, features,numeric ,character or string sequences etc,datset may be of any type like attack groups,images,SMS collection,Audio or video,Biometric datasets, etc.Various organizations provides dataset for research purpose based on their types and applications Here comparative study of different dataset has been presented for research purpose or for analyzing system behavior.

Related work: Another approach, payload based anomaly detector (PAYL) extends the work in [5] by utilizing a different statistical model, a full byte distribution and the use of clustering [3] which they utilize to detect worms. Though they state that their approach could detect other types of attacks, it was designed specifically to detect worms. They tested their approach with other web attacks using the 1999 DARPA dataset. They also used datasets from their university web server to detect Code Red and a buffer overflow attack. In terms of their character distribution these two attacks are vastly different to normal requests. In this paper we focus on the detection of more subtle attacks [1,2].

Propose Work: Here in the given work various dataset approaches has been given as every system should be tested on universal datasets

to get desired result and system parameters evaluation. Below datasets are shown.

1) KDD CUP- DARPA DATASET:

It Contains attack cases and attack varieties, It contains TCP –dump that is data available on network for analyzing network parameters, The dataset has different representation techniques, here datasets are represented by attack features.

1) DSHIELD DATASET:-

It contains different attacks and classes of IP address, ports, URL's, which helps in determining attack types and causes. The available List helps in creating recommendation systems for generating alert to user .

1)SQL-INJECTION DATASET(Query String)- User created dataset:-

It contains two classes of Sql-queries like:

a) Original query(Safe query)- 'O' class:

Select * from student where username=" ";

b) Suspicious query (Unsafe)- 'S' Class:

Select * from student where username=" " OR 1=1--;

These 2 different classes of query strings are used to classify the Input entered by user.

1)UCI MACHINE LEARNING REPOSITORY:-

It contains 225 data sets for machine learning service approach.

1)CAIDA(Cooperative Association of Internet Data Analysis):

It contains Dos attack datasets.

1)KDNUGGETS DATASET:

It is a collection of different datasets(Ex. KDD CUP,UCI machine Learning),It contains different datasets application for different techniques like Artificial intelligence vfield,machine learning fields, attacks categorizations.

1) **CLUEWEB 09 DATASET**:

It Contains web application datasets. It represents web design and architectural datasets.

2) **SMS SPAM COLLECTION DATASET**:

It contains SMS labeled messages of public set collected for mobile phone spam research.

2) MAIL SERVER DATABASE FOR RESEARCH:

It contains data problem in machine learning for collaborative prediction and ranking with non random missing data.

2) MULTIMEDIA WEBSITES DATASET:

It is used for generating directed graph techniques from normal crawl process of videos' contains records of multimedia taken from different applications and sites.

2)EMAIL DATASET :

It contains data or emails of different users, collected from different mail-server.whcih is used for creating analyzing, mail features and its contents, it may be attacks, hidden code etc.

2) **GOOGLE BOOKS N-GRAM DATASET:**

It is used for crawling techniques and for web page links. Here different pages links and their response and behavior are shown for analyzing searching and indexing technique.

2) IMAGE NET(Recognition dataset):

It uses for image ,face, speech recognition techniques. It represents techniques used for creating recognition system applied at different system for getting secure input and to check authenticity.

2) LENDING CLUB LOAN DATASET:

Use for Loan Related research. It is used for creating bank related research for generating schema of popularity and bank policies amendments.

2) MOVIELENS DATASETS/GROUPLENS RESEARCH DATASET:

Use for Movies and image type research. Different movies and clips data which are used to analyze movie and clip size and behavior at different systems.

2)CENSUS BUREAU DATASET:

Contains population and economic indication. And use for population counting and behavior type research.

2)**PUBLIC DATASETS ON AMAZON S3 WEB SERVICES:**

Contains centralized repository of public data used in cloud based applications.

2)QUANTUM CHAOTIC-FACEBOOK 100 DATASET:

Contains datasets of friends and their relationships, use for analyzing social networking site behavior.

2)**WORDNET DATASET:**

Contains large lexical database of English nouns, verbs, adjectives and adverbs are grouped into sets, and contains synonyms and concepts.

2)MILLION SONG RESEARCHSCALING MIR RESEACR DATASET:

Contains song datasets, their metadata and features.

2)DATA/WORLD BANK DATASET:

Contains dataset of countries , years and indicators ,used for inconsistencies and apparent data error reports.

2)EMNLP-2011:

Contains machine translation statistical dataset.

2)ALIGNED HANSARDS(PARLIAMENT OF CANADA):

Contains official records of parliament ,provided for research. Dataset based for the study of political behavior.

2)CRCNS DATASET(COLLABORATIVE RESEARCH IN COMPUTATIONAL NEUROSCIENCE-DATA SHARING):

Use in neuroscience research. brain mapping systems, neural networks are the field ,where the propose dataset has vast applications.

2)**UNIGENE DATASET(BIO DATASET):**

A biological dataset used for the study of genetic features.

2)GENE EXPRESSION OMNIBUS(GEO) DATASET:

Contains repository of public functional genomics.

2)**SOCIAL SCIENCE DATASET:**

Contains geographical or historical type data for research purposes.

2)PEW RESEARCH CENTER 'S AND AMERICAN LIFE PROJECT DATASET:

Use for creating research on Americans life styles.

FLICKS PERSONAL TAXONOMIES DATASET:

Use for social networking research purposes. Contains weblog and social media dataset, used for creating patterns in web attacks.

Result: The dataset shown above contain different features and application. Each dataset has its own application and analyzing pattern. These are provided by standard organization, they are timely or even daily updated by the dataset providers, Purpose of using standard dataset are, they provides standard result which are applicable and acceptable at different varieties of hardware or software platforms.

Conclusion:

A brief study of different datasets has been creating in the propose paper, here, it is analyzed that, at much extends system performance depends on the type and feature s of dataset taken for their study.

References:

[1] .Mehdi Kiani, Andrew Clark and George Mohay,"Evaluation of Anomaly Based Character Distribution Models in the Detection of SQL Injection Attacks",IEEE 2008 .

[2]. Huang, F. Yu, C. Hang, C. H. Tsai, D. T. Lee, and S. Y. Kuo. —Securing Web Application Code by Static Analysis and Runtime Protection‖, *Proc* 12th International World Wide Web Conference (WWW 04), May 2004.

[3]. SQL Injection Attack Examples based on the Taxonomy of Orso et al.

[4]. G.T. Buehrer, B.W.Weide and P.A..G.Sivilotti, "Using Parse tree validation to prevent SQL Injection attacks", In proc. Of the 5th International Workshop on Software Engineering and Middleware (SEM '056), pp.106-113, Sep. 2005.

[5] www.google.com/research-datasets.