IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Role Of Pattern Recognition In Computer Virus Detection

Ankur Singh Bist
Govind Ballabh Pant University of Agriculture And Technology, Panthnagar, India
ankur1990bist@gmail.com

Abstract

Pattern recognition is a well known technique used for solving various problems . Computer virus problem is a big issue that is needed to be handled .In this paper we will analyse the various aspects of computer virus classification and detection with the help of pattern recognition .

Keywords: Pattern , Signature .

Introduction

Pattern recognition is a technique of detecting pattern .This technique is widely used in various applications like handwriting matching , fingerprint matching and in various medical applications . Yas Abbas Alsultanny, Musbah M. Aqel in his paper Pattern recognition using multilayer neural-genetic algorithm explained about application of pattern recognition with decision theoretical model of pattern recognition system.

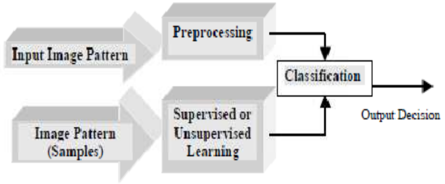


Figure . Decision theoretical pattern recognition system[1]

Another activities carried our by the process of pattern recognition involves:-
1. Collect data
2. Select features
3. Select model
4. Train classifier
5. Evaluate classifier

Data collection involves the primary step of collecting all the sample data that is to be tested or simply called input data . Feature selection involves the step of determining various similar features and also those features that are responsible for discrimination with this it is also essential to determine robust features . There are different types of model for selection for particular domain such as[2]-------

1. Templates
2. Statistical
3. Syntactic
4. Neural
5. Hybrid

Training involves the process of learning that involves various subtypes in itself :-
1. Supervised learning
2. Unsupervised learning
3. Reinforced Learning

Evaluate classifier involves the process of deciding the performance issues that involves the efficiency issues associated . The following object diagram of pattern recognition explains various activities involved in pattern recognition

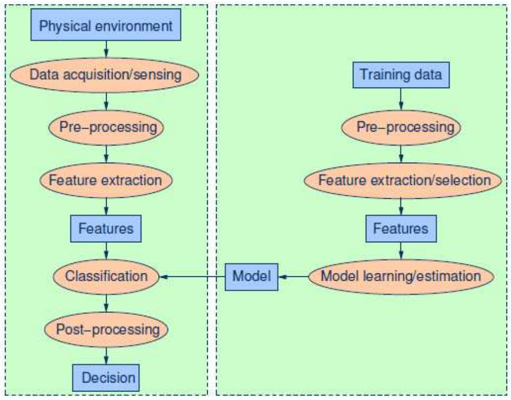
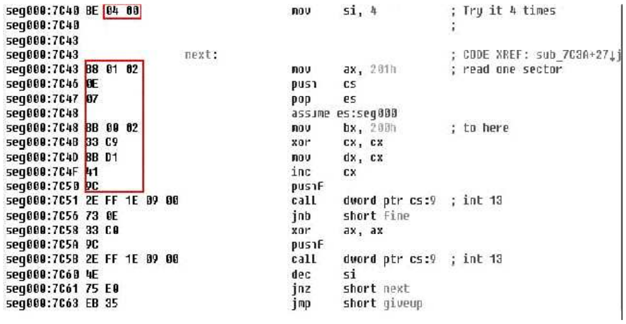


Figure .Object diagram of pattern recognition system[2]

Pattern Recognition And Computer Virus Static Detection

Computer virus detection is a very important issue in term of computer security measures . Computer virus detection firstly involves to classify viruses from normal file . One simple technique is called the signature detection .



```
seg000:7C40 BE 04 00                    mov     si, 4           ; Try it 4 times
seg000:7C40                                                     ;
seg000:7C43
seg000:7C43             next:                                   ; CODE XREF: sub_7C3A+27↓j
seg000:7C43 B8 01 02                    mov     ax, 201h        ; read one sector
seg000:7C46 0E                          push    cs
seg000:7C47 07                          pop     es
seg000:7C48                             assume es:seg000
seg000:7C48 BB 00 02                    mov     bx, 200h        ; to here
seg000:7C4B 33 C9                       xor     cx, cx
seg000:7C4D 8B D1                       mov     dx, cx
seg000:7C4F 41                          inc     cx
seg000:7C50 9C                          pushf
seg000:7C51 2E FF 1E 09 00              call    dword ptr cs:9  ; int 13
seg000:7C56 73 0E                       jnb     short fine
seg000:7C58 33 C0                       xor     ax, ax
seg000:7C5A 9C                          pushf
seg000:7C5B 2E FF 1E 09 00              call    dword ptr cs:9  ; int 13
seg000:7C60 4E                          dec     si
seg000:7C61 75 E0                       jnz     short next
seg000:7C63 EB 35                       jmp     short giveup
```

Figure : Stoned virus showing the search signature 0400 B801 020E 07BB 0002 33C9 8BD1 419C (Szor, 2005c)[4]

The most famous technique in anti-virus scanners is signature based detection. It is not as much effective as other techniques available but it can be performed more quickly. This technique involves the process of extracting a unique sequence of bits from a known virus code and this sample is subsequently used to match against while scanning for existence of the virus. Attention has to be taken when choosing the bit sequence to reduce the probability of the number of false positives and false negative at the same time match the virus and possible variants. Statistical techniques is one of the efficient techniques , used to extract these patterns. Signature based technique involves the process of pattern recognition that is matching the signature of detected file signature with the signatures of infected file available in detection tool database but the problem is that the signature of new born viruses can not be available in database so this process that includes pattern matching is not sufficient and it shows that it is required to evolve better techniques .

Second generation signature based detectors use more advanced techniques that involves [3]:-

1. Smart scanning -- ignoring nop instructions
2. Wildcards----allowing skipping of bytes and byte ranges
3. generic matching ----using a single string to potentially match a family of viruses
4. Near exact identification---using two search strings instead of one
5. Checksum

Pattern Recognition And Computer Virus Dynamic Detection

For detecting computer virus of polymorphic and metamorphic nature it is require to adopt some better techniques. There are various methods based on pattern recognition design approaches of Collect data , Select features , Select model , Train classifier and Evaluate classifier . One of the technique using this is Hidden Markov Models that are well-known for their use in speech recognition . other applications include protein sequences modeling for protein families and patterns in RNA splice junctions . Hidden Markov Models are used for detecting metamorphic viruses produced good results , Profile Hidden Markov Models also used in detecting metamorphic variants of viruses . Profile Hidden Markov Models [3] are widely used in Bioinformatics for finding distantly-related sequences of a protein sequence family . This usage of profile hidden Markov model inspired the researchers to use this method for virus detection specially for metamorphic viruses. Various data mining techniques includes the concept of pattern recognition are used for detecting the computer viruses. Techniques like Pairwise alignment that are used for similarity analysis of proteins includes pattern recognition method , now used for detecting computer viruses. Different tools are also available to study the analysis of pattern recognition and to implement it according to user problem domain . One of the most common tool is matlab and hidden Markov model toolbox is also available for it . Data mining tool weka can also be used in same domain. There are some essential factors require for pattern recognition tool like for the automatic recognition of the classes of objects, first some measurements have to be collected, e.g. using sensors, then they have to be represented, e.g. in a feature space and after some possible feature reduction steps they can be finally mapped by a classifier on the set of class labels. Between the initial representation in the feature space and this final mapping on the set of class labels the representation may be changed several times: simplified feature spaces (feature selection), normalization of features (e.g. by scaling), linear or nonlinear mappings (feature extraction), classification by a possible set of classifiers, combining classifiers and the final labeling. In each of these steps the data is transformed by some mapping.

Two basic concepts of PRTools are defined[5]:

Datasets: matrices in which the rows represent the objects and the columns the features, class memberships, or other fixed sets of properties (e.g. distances to a fixed set of other objects). In PRTools4.1 an extension of the dataset concept has been defined: data files. These refer to datasets to be created from directories of files.

Mappings: transformations operating on datasets. As pattern recognition has two stages, training and execution, mappings have also two types: untrained and trained.

An untrained mapping refers just to the concept of a method, e.g. forward feature selection, PCA, or Fisher's linear discriminant. It may have some

parameters that are needed for training, e.g. the desired number of features or some regularization parameters. If an untrained mapping is applied to a dataset it will be trained (training).
A trained mapping is specific for the training set used to train the mapping and more details about prtools can be taken from reference 5 .

Conclusion

In this paper firstly we take a look on various aspects of pattern recognition including various application areas of pattern recognition . Various approaches of pattern recognition taken by different authors are taken . Finally the main issue of computer viruses in two phases are discussed including static and dynamic detection which includes various approaches inspired by pattern recognition like profile hidden Markov model . At last various tool are discussed to analyse the various issues of pattern recognition .

References

[1] Yas Abbas Alsultanny, Musbah M. Aqel ," Pattern recognition using multilayer neural-genetic algorithm"
[2] Selim Aksoy ," introduction to pattern recognition" CS 551, Fall 2012 .
[3] "DETECTING METAMORPHIC VIRUSES USING PROFILE HIDDEN MARKOV MODELS " A Project Report Presented to The Faculty of the Department of Computer Science San Jose State University.
[4] SYAHRIZAL AZMIR BIN MD. SHARIF ," ANALYSIS AND EFFECTIVENESS OF SIGNATURE BASED IN DETECTING METAMORPHIC VIRUS"
[5] http://www.prtools.org/files/PRTools4.1.pdf