# Requirements to Build a System that Uses Machine Learning Based Approach for Analysis of Forensic Data

Shruti B. Yagnik
*Department of Computer Science,*
*IT Systems and Network Security,*
*Gujarat Technological University,*
*India*

***Abstract*- Cyber Forensic Investigation paradigm is laborious and requires significant expertise on the part of the investigators. Computer Intelligence and Cyber Forensics together is a match-winning integration for making investigation tasks easy, accurate and less laborious. This research paper focuses on the requirements to build a system that is automatic and highly efficient and assists in Forensic Investigations by binding Computer Intelligence and Machine Learning Technology to Computer Forensic Framework. It consists of building a Learning Model which learns efficiently from Data Sets through Intelligent Algorithms and various Learning Approaches. This training provided, intelligently classifies accordingly all the further inputs given from real time cyber-crime data. We will acquire the basic knowledge of how to develop such systems, which work on machine learning and help in forensic investigations.**

***Keywords*- Computer Forensic Framework, Cyber-Crime, Machine Learning, Cyber Forensics, Computer Intelligence, Cyber Forensic Investigations, Learning Model, Data Sets, Intelligent Algorithms, Learning Approaches.**

## I. INTRODUCTION

We need to build computer systems that automatically improve with experience and help in forensic investigations by learning and doing classifications according to pattern recognition and data mining. It should label data and classify them accordingly.

### *A. Machine Learning*

Machine Learning is a branch of Artificial Intelligence concerned with developing computer algorithms that can learn rules from data, Adapt to changes and can improve performance with experience [1]. Machine learning is the science of getting computers to act without being explicitly programmed [2]. A major focus of machine learning research is the design of algorithms that recognize complex patterns and make intelligent decisions based on input data. It is concerned with the design and development of algorithms that take as input empirical data, such as that from sensors or databases, and yield patterns or predictions thought to be features of the underlying mechanism that generated the data [2].

In general, a machine learning task can be defined formally in terms of three elements, viz. the learning experience E, the tasks T and the performance element P. Machine Learning can be defined as a learning task more precisely as, A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. This representation of a machine learning task clearly defines the requirements. It gives an idea of what the machine learning problem is, and what are its learning goals. It also states how these goals can be measured so that the effectiveness of the task can be decided [3].

### *B. Cyber Forensics*

"The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitation or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations" [4]. It attempts to reconstruct events, focusing on the computer based conduct of an individual or group of individuals.

Evidences includes everything that is used to determine or demonstrate the truth of an assertion. Physical evidence cannot be wrong, it cannot perjure itself, and it cannot be wholly absent. Only human failure to find it, study and understand it, can diminish its value" [5].

## II. Requirement Discovery Using Machine Learning

Before we create a system, we must know the basics as to why we are developing the system, its goals, the various requirements, scenarios that result into such investigations, the input outputs of such systems, and methods to train such system. We should also be clear regarding the algorithm to be used and the approach to be followed in order to design an appropriate system that will work perfectly well for the investigation of crime.

### A. Goals

The main goal of Machine Learning in Computer Forensics is to perform investigation on a computer and find who was responsible for it. Investigators follow a standard set of procedures using Machine Learning Tools, Algorithms and Approaches on a physically isolated computer in question to make sure it cannot be accidentally contaminated, and collect evidences by making a digital copy of the hard drive. Once evidence is taken, it is locked in a safe or other secure storage facility to maintain its pristine condition. All investigations are done on the digital copy [1].

### B. Scenarios

Computer Forensics is a practice of investigating storage media for the purpose of discovering and analysing available, deleted, or hidden data that may serve as evidences in a legal matters. It can also be used to uncover evidences in cases like Copyright infringement, Industrial espionage, Piracy, Sexual harassment, Child pornography, Theft of intellectual property, unauthorized access to confidential information, Destruction of information, Fraud, Illegal duplication of software.

Retrieve deleted evidence - operating systems utilizes a directory that contains the name and placement of each file on the drive. When a file is deleted, file status marker is set to show that the file has been deleted. A disk status marker is set to show that the space is now available for another use. The user cannot see the file listed in the directory but the file actually exists. This newly available space is called free or unallocated space and until the free space is overwritten by another file, the forensic specialist can retrieve the file in its entirety. If data is overwritten only then the old data is not retrievable.

## III. Basics for Building a Learning System

There are some of the basics that developers should be aware of in order to build up an effective and efficient Machine Learning System. They must develop according to the following approaches [12].

1. *Learner:* Who or what is doing the learning, an algorithm or a computer program.
2. *Domain:* What is being learned, a function, or a concept.
3. *Representation:* The way the objects to be learned are represented the way they are to be represented by the computer program.
4. *Algorithmic technology:* The algorithmic framework to be used.
5. *Information source:* The information (training data) the program uses for learning. This could have different forms: positive and negative examples (called labelled examples), answers to queries, feedback from certain actions, and so on. An information source may be noisy, i.e. the training data may have errors. Examples may be clustered before use in training a program.
6. *Training scenario:* The description of the learning process. In an on-line learning scenario, the program is given examples one by one, and it recalculates its hypothesis of what it learns after each example. Examples may be drawn from a random source, according to some known or unknown probability distribution.
7. *Prior knowledge:* What is known in advance about the domain, e.g. about specific properties (mathematical or otherwise) of the concepts to be learned. This might help to limit the class of hypotheses that the program needs to consider during the learning, and thus to limit its 'uncertainty' about the unknown object it learns and to converge faster.
8. *Success criteria:* The criteria for successful learning. Depending on the goal of the learning program, the program should befit for its task. If the program is used e.g. in safety-critical environments, it must have reached sufficient accuracy in the training phase so it can decide or predict reliably during operation.
9. *Performance:* The amount of time, space, resources and computational power needed in order to learn a certain task, and also the quality (accuracy) reached in the process.

## IV. The System Learning Model

A

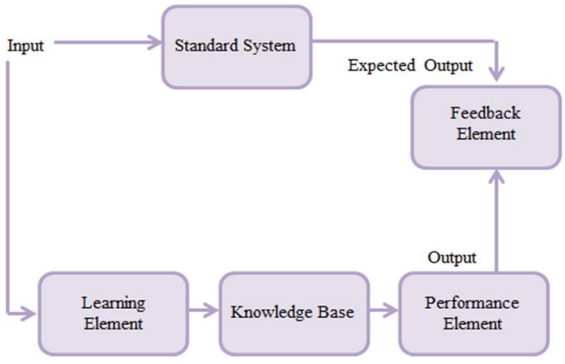


Figure 1: Learning System Model [6]

machine learning system usually starts with some knowledge and a corresponding knowledge organization so that it can interpret, analyse, and test the knowledge

acquired. The figure shown above is a typical learning system model. It consists of the following components [6].

*Learning element*: It receives and processes the input obtained from a person (teacher), from reference material like magazines, journals, etc., or from the environment at large.

*Knowledge base*: This is somewhat similar to the database. Initially it may contain some basic knowledge. Thereafter it receives more knowledge which may be new and so be added as it is or it may replace the existing knowledge.

*Performance element*: It uses the updated knowledge base to perform some tasks or solves some problems and produces the corresponding output.

*Feedback element*: It is receiving the two inputs, one from learning element and one from standard (or idealized) system. This is to identify the differences between the two inputs. The feedback is used to determine what should be done in order to produce the correct output.

*Standard system (Idealized System)*: It is a trained person or a computer program that is able to produce the correct output. In order to check whether the machine learning system has learned well, the same input is given to the standard system. The outputs of standard system and that of performance element are given as inputs to the feedback element for the comparison.

## V. Input Output Functions of a Machine Learning System

While designing a system we need to keep into consideration the various inputs and outputs that we give to that system. This is because we need to accordingly design the storage facility for the system. We need to make better retention and transfer facilities for that system. All this depends on the input you provide and the output that you get from the system. We select hypothesis based on a training set, of m input vector examples [7]. Many important details depend on the nature of the assumptions made about all of these entities [7]:

### *A. Type of Learning*

Machine learning Types can be organized into taxonomy based on the desired outcome of the algorithm.

- Supervised learning generates a function that maps inputs to desired outputs (also called labels, because they are often provided by human experts labelling the training examples).
- Unsupervised learning models a set of inputs, like clustering. They are unlabelled data.
- Semi-supervised learning combines both labelled and unlabelled examples to generate an appropriate function or classifier.
- Reinforcement learning learns how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback in the form of rewards that guides the learning algorithm.
- Transduction learning tries to predict new outputs on specific and fixed (test) cases from observed, specific (training) cases.
- Learning to learn learns its own inductive bias based on previous experience.

### *B. Input Vectors*

Because machine learning methods derive from so many different traditions, its terminology is rife with synonyms, and we will be using most of them in this book. For example, the input vector is called by a variety of names. Some of these are: input vector, pattern vector, feature vector, sample, example, and instance. The components, xi, of the input vectors are variously called features, attributes, input variables, and components. The values of the components can be of three main types. They might be real-valued numbers, discrete-valued numbers, or categorical values. As an example illustrating categorical values, information about a student might be represented by the values of the attributes class, major, sex, and adviser. A particular student would then be represented by a vector such as: (sophomore, history, male, Higgins). It is also possible to represent the input in unordered form by listing the names of the attributes together with their values. As an example of an attribute-value representation, we might have: (major: history, sex: male, class: sophomore, adviser: Higgins, age: 19).

### *C. Outputs*

The output may be a real number, in which case the process embodying the function, h, is called a function estimator, and the output is called an output value or estimate. Alternatively, the output may be a categorical value, in which case the process embodying that is variously called a classifier, a recognizer, or a categorizer and the output itself is called a label, a class, a category, or a decision. Classifiers have application in a number of recognition problems, for example in the recognition of hand-printed characters. The input in that case is some suitable representation of the printed character, and the classifier maps this input into one of, say, 64 categories.

Vector-valued outputs are also possible with components being real numbers or categorical values. An important special case is that of Boolean output values. In that case a training pattern having value 1 is called a positive instance, and a training sample having value 0 is called a negative instance. When the input is also Boolean, the classifier implements a Boolean function. We study the Boolean case in some detail because it allows us to make important general points in simplified setting. Learning a Boolean

function is sometimes called concept learning, and the function is called a concept.

*D. Training Regimes*

There are several ways in which a training set can be used to produce hypothesized function. In the batch method, the entire training set is available and used all at once to compute the function h. A variation of this method uses the entire training set to modify a current hypothesis iteratively until an acceptable hypothesis is obtained. By contrast, in the incremental method, we select one member at a time from the training set and use this instance alone to modify a current hypothesis. Then another member of the training set is selected, and so on. The selection method can be random (with replacement) or it can cycle through the training set iteratively. If the entire training set becomes available one member at a time, then we might also use an incremental method of selecting and using training set members as they arrive. Using the training set members as they become available is called an online method.

*E. Noise*

Sometimes the vectors in the training set are corrupted by noise. There are two kinds of noise. Class noise randomly alters the value of the function; attribute noise randomly alters the values of the components of the input vector. In either case, it would be inappropriate to insist that the hypothesized function agree precisely with the values of the samples in the training set.

*F. Performance Evaluation*

Even though there is no correct answer in inductive learning, it is important to have methods to evaluate the result of learning. In supervised learning the induced function is usually evaluated on a separate set of inputs and function values for them called the testing set. A hypothesized function is said to generalize when it guesses well on the testing set. Both mean-squared-error and the total number of errors are common measures

## VI. General Requirements for Machine Learning System

The main ways of how to transfer the data from one section to another into the system is taken into consideration. We need to facilitate the storage of data and its transfer. This is an important factor to be taken into consideration [13].

*A. Form of knowledge retention*

Knowledge learnt by machines can be stored in within a Machine Learning System [8]. To retain task knowledge, save the respective training examples. You can also store or model search parameters such as the learning rate in neural networks. An advantage of retaining actual training examples is the accuracy and purity of the knowledge. The advantages of retaining representational knowledge is its compact size relative to the space required for the original training examples and its ability to generalize beyond those examples. Disadvantages of retaining knowledge are the large amount of storage space that it requires and difficulties in using such knowledge during future learning.

*1) Requirements for Long-term Retention of Learned Knowledge*

*a. Effective and Efficient retention*

A Machine Learning System should resist the introduction and accumulation of domain knowledge error. Only hypotheses with an acceptable level of generalization accuracy should be retained else, once saved in long-term memory, the error from a hypothesis may be transferred to future hypotheses. Similarly, the process of retaining a new hypothesis should not reduce its accuracy or the accuracy of prior hypotheses existing in long-term memory. The integration of new task knowledge should increase the accuracy of related prior knowledge. Machine Learning System should be efficient in its use of long-term memory and also should be efficient in storage. In particular, the system should make use of memory resources such that the duplication of information is minimized.

*b. Effective and Efficient indexing*

A Machine Learning System must be capable of selecting the appropriate prior knowledge for transfer during short-term learning. This requires that it should be capable of indexing into long-term memory, task knowledge that is most related to the primary task. Typically, primary task knowledge will arrive in the form of training examples and no representational knowledge will be provided. The system must also make the selection of related knowledge as rapid as possible. Preferably, the computational time for indexing into domain knowledge should be no worse than polynomial in the number of tasks having been stored.

*c. Meta-knowledge of the task domain*

In most cases, it will be necessary for such systems to determine and retain meta-knowledge of the task domain. It is necessary to estimate probability distribution over input space to manufacture appropriate functional examples from retained task representation [9]. Alternatively, it may be necessary to retain characteristics of the learning process for each task.

*B. Form of knowledge transfer*

This requirement deals with the form in which knowledge retained is transferred. For example, the retained hypothesis representation for a learned task can be used to generate functional knowledge in the form of training examples. Representational transfer involves direct or indirect assignment of known task representation to the model of a new target (or primary) task [8]. In this way the learning system is initialized in favour of a particular region of hypothesis space of the modelling system [9]. Representational transfer often results in substantially reduced training time with no loss in the generalization performance of the resulting hypotheses. In contrast to representational transfer, functional transfer employs the use of implicit pressures from training examples of related tasks, the parallel learning of related tasks constrained to use a common internal representation or the use of historical training information from related tasks.

*2) Requirements for Short-term Learning with Inductive Transfer*

*a. Effective and Efficient learning*

The inductive transfer from long-term memory should never decrease the generalization performance of a hypothesis developed by Machine Learning System. The system should produce hypothesis for the primary task that meets or exceeds the generalization performance of that developed strictly from the training examples. There is evidence that the functional form of knowledge transfer somewhat surpasses that of representation transfer in its ability to produce more accurate hypotheses [10, 11]. Starting from a prior representation can limit the development of novel representation required by the hypothesis for the primary task. In terms of neural networks this representational barrier manifests itself in terms of local minimum. Inductive transfer from long-term memory should not increase the computational time for developing a hypothesis for the primary task as compared to using only the training examples. In fact, inductive transfer should reduce training time. In practice this reduction is rarely observed because of the computation required to index into prior domain knowledge. In terms of memory (space), there will typically be an increase in complexity as prior domain knowledge must be used during the learning of the new task. Our research has shown that a representational form of knowledge transfer will be more efficient than a functional form [10].

*b. Transfer versus training examples*

A Machine Learning System must take into consideration the estimated sample complexity and number of available examples for the primary task and the generalization accuracy and relatedness of retained knowledge in long-term memory. During the process of inductive transfer the system must weigh the relevance and accuracy of retained knowledge alongside that of the information resident in the training examples.

*C. Input and output type, complexity and cardinality*

The output representation of a system capable of retaining and transferring knowledge should not be constrained to a particular data type. A Machine Learning System should be capable of predicting class categories and real-value outputs including scalar values as well as vectors. It should also be capable of dealing with its environment over a lifetime with a fixed number of inputs and outputs for the tasks under study. Certain inputs or outputs might go unused for many tasks of a domain early in the learning system's lifetime only to be used quite frequently later in life. This ensures a consistent interface with the environment and with other entities such as a software agent, an application program or a human user.

*D. Scalability*

A Machine Learning System must be capable of scaling up to large numbers of inputs, outputs, training examples and learning tasks. Preferably, both the space and time complexity of the learning system grows extensively in all three of these factors.

*E. Performance of a Machine Learning System*

The system should be developed keeping in mind the performance. The system developed should provide high performance along with accurate results.

*1) Types of training provided*

There are several ways in which a training set can be used to produce a hypothesized function. In the batch method, the entire training set is available and used all at once to compute the function h. A variation of this method uses the entire training set to modify a current hypothesis iteratively until an acceptable hypothesis is obtained. By contrast, in the incremental method, we select one member at a time from the training set and use this instance alone to modify a current hypothesis. Then another member of the training set is selected, and so on. The selection method can be random (with replacement) or it can cycle through the training set iteratively. If the entire training set becomes available one member at a time, then we might also use an incremental method of selecting and using training set members as they arrive. Using the training set members as they become available is called an online method.

*2) The form and extent of any initial background knowledge*

The initial background information should be stored in the system. Background knowledge can be used to make learning more efficient by reducing the search space and faster results.

*3) The type of feedback provided*

The feedback provided also makes a great deal of impact on the performance of the system. The feedback system takes input from two places. One is the computer, programmed with logic and giving the cyber forensic investigation result. Whereas the other is the idealized system. The difference between these two results is given by the feedback system. This should be efficient so as to eradicate the difference in the results of both these systems.

*4) The learning algorithms used.*

The success of machine learning system also depends on the algorithms. It depends on the type of the application that you are making. Here we focus on cyber forensic data analysis. We should select the algorithm in such a manner that it gives the most accurate results in all environments and also is the optimal way to perform that type of investigations. It should require the minimum of time, money and resources.

## VII. CONCLUSION

The purpose of this paper was to learn that Machine Learning along with its algorithms and Approaches should be integrated into every phase of the Digital Investigation Process to analyse and reconstruct the crime scene. The investigation procedure becomes faster, less tedious and laborious through this perfect combination.

## ACKNOWLEDGEMENT

I am grateful to my Guide Mr Kaushal Bhavsar for his patience, guidance and constant encouragement. I appreciate the thoughtful contributions that he has made to my efforts. It has been a great privilege to work under him. Without his guidance and persistent help this research and survey would not have been possible. I would like to thank Mr Jatin Raval, who with his valuable feedback and guidance made it possible to work on this topic. I also wish to thank him for his comments on various aspects of this draft and for his valuable input in this work. I greatly owe to my friends and colleagues for their help, suggestions and actively participating in my efforts of making this document. I have learnt a lot from them academically and personally. I heartily acknowledge the support and help of several people who have been instrumental in some way or the other for making this survey a pleasant reality.

## REFERENCES

[1] Dalal Alrajeh, Requirements Discovery using Machine Learning, Department of Computing, Imperial College London, 2001.

[2] Wikipedia (2012, Sept 12) *Machine Learning* [Online]. Available:http://en.wikipedia.org /wiki/Machine_learning

[3] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997

[4] Brian Carrier, Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers, *International Journal of Evidence*, Winter 2003, Volume 1, Issue 4.

[5] *Crime investigation: physical evidence and the police laboratory*. Interscience Publishers, Inc.: New York, 1953

[6] *Articles on Artificial Intelligence* [Online]. Available: http://intelligence.worldofcomputing.net/machine-learning/machine-learning-overview.html

[7] Nils J. Nilsson, Introduction to machine learning: an early draft of a proposed textbook, Robotics Laboratory, Department of Computer Science, Stanford University, Stanford, November 3, 1998.

[8] Daniel L. Silver and Robert E. Mercer. The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness. *Connection Science Special Issue: Transfer in Inductive Systems*, 8(2):277–294, 1996.

[9] Daniel L. Silver and Robert E. Mercer. The task rehearsal method of life-long learning: Overcoming impoverished data. *Advances in Artificial Intelligence, 15th Conference of the Canadian Society for Computational Studies of Intelligence (AI'2002)*, pages 90–101, 2002.

[10] Daniel L. Silver and Peter McCracken, Selective transfer of task knowledge using stochastic noise. In Yang Xiang and Brahim Chaib-draa, editors, *Advances in Artificial Intelligence, 16th Conference of the Canadian Society for Computational Studies of Intelligence (AI'2003)*, pages 190–205. Springer-Verlag, 2003.

[11] Richard A. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[12] Jan van Leeuwen. Approaches in machine learning. Institute of Information and Computing Sciences, Utrecht University, Padualaan 14, 3584 CH Utrecht, the Netherlands.

[13] Daniel L. Silver and Ryan Poirier, Requirements for Machine Lifelong Learning, Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada B4P