International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Attacking & Preventing an AI Based Disease Identification System

Prasanna Shivaji Shinde[1], Avinash Shrivas[2]

[1, 2] Mumbai University, Vidyalankar Institute of Technology, Wadala, Mumbai

Abstract: ***Artificial intelligence in medical (AIM) has reached a period of adolescence in which interactions with outside world are not only natural but mandatory. Furthermore, an adequate appraisal of AIM research requires an understanding of the research motivations, the complexity of problems and a suitable definition of criteria for judging the field's success. In this proposed system user can communicate to the system for retrieving information about diseases by giving symptoms as input. Also, an algorithm independent approach for mounting poisoning attack across a dataset of system is being stated along with the countermeasure for attack which specifies how long the dataset of system is secure.***

Keywords: AIM, Chatbot, Poisoning Attack, machine learning, NLP

1. Introduction

Much of our work on language, speech, translation, and visual processing relies on machine learning and AI. In all of those tasks and many others, we gather large volumes of direct and indirect evidence of relationship of interest, and we apply learning algorithm to generalize from that evidence to new cases of interest. Machine learning algorithms are being used in critical applications where they drive decisions with large personal, organizational, or societal impact. These applications include healthcare, network intrusion detection systems (IDSs), spam and fraud detection, phishing detection, political decision making, and financial engineering.

The branch of computer science concerned with making computers behave like humans means interaction between computers and human languages. The process of a computer extracting meaningful information from natural language input and producing natural language output. There are many sites which provide the facility of AI in computer. And ALICE is highly recommended chat Robot. Currently the most famous and intuitive chat service is provided by them. We intend to make such AI application that it is useful for patients. User can communicate to the system for retrieving information about diseases by giving symptoms as input. Also medical related information will be provided by our bot.

Two main categories of security attacks on machine learning have been considered in the literature: exploratory and causative. Exploratory attacks exploit existing vulnerabilities without altering the training process. On the other hand, causative attacks alter the training process, typically by modifying the training dataset. Poisoning attacks are a class of causative attacks in which carefully-crafted malicious instances are added to the training dataset, leaving the rest of the dataset intact.

New results indicate that it may be easier than we thought to provide data to a learning program that causes it to learn the wrong things. The key idea is that, given machine intelligence, the trick to defeating it is to feed it the wrong data. Security experts call the idea of breaking a system by feeding it the wrong data a poison attack. The approach also provides the countermeasure for measuring the defending power of the application dataset against the poisoning attack.

In the healthcare system, hindrance of a diagnosis may have life threatening consequences and could cause distrust. On the other hand, not only may a false diagnosis prompt users to distrust the machine learning algorithm and even abandon the entire system but also such a false positive classification may cause patient distress.

It could be possible to direct the induced errors so as to product particular types of error. For example, a spammer could send some poisoned data so as to evade detection in the future. The biggest practical difficult in using such methods is that, in most cases, the attacker doesn't control the labeling of the data points - i.e. spam or not spam - used in the training. A custom solution would have to be designed to compromise the labeling algorithm.

These attacks may be applied to both fixed and evolving datasets. They can be applied even when only statistics of the training dataset are available or, in some cases, even without access to the training dataset. Finally, we present countermeasures against the proposed generic attacks that are based on tracking and detecting deviations in various accuracy metrics, and benchmark their effectiveness.

2. Previous Work

Depending upon the research based on this system we have following topics available, which state methodologies for implementing chatbot and attacking and securing the attack on machine learning algorithm used for creating medical dataset. These applications are based on Artificial Intelligence.

2.1 Language Translator

This is one of the most important applications of Natural Language Processing. Translation of a sentence from one language to another, retaining the meaning, is a difficult task. A lot of research has been done on this now in different parts of the world.

2.2 Existor: Chatbot named Evie

There are no limits. Artificial Intelligence is communication. Natural language is universal. Evie was created by Existor. Evie is an Electronic Virtual Interactive Entity. This bot is capable of communicating with human beings in very normal way and also on general topics. Also it had now learned to understand smiley's and meaning of other symbols. In early September 2011, Existor's chatbot Cleverbot, recently took the Turing Test at the Techniche computer festival in Guwahati - and was judged to be 59% human! The humans were only judged to be 63% human, so Cleverbot has arguably passed the test.

2.3 Dr. Romulon: chatbot

This chatbot is based on the ALICE artificial intelligence chat platform. Unlike the Eliza chatbot, this bot has a larger set of responses. But the Doctor does have a very poor memory. Also this chatbot is unable to learn from experiences and also from past conversations.

2.4 Eliza : Chatbot (Psychotherapist)

Eliza is one of the oldest chat programs around. She is the classic chat bot, born 46 years ago in 1966. Eliza is a conversation simulation. But more so, she emulates a psychotherapist in the way she asks and answers your statements. Also this chatbot is unable to learn from experiences and also from past conversations.

2.5 A Basic Program for Clinical Problem-Solving

Doctors use a combination of a patient's case history and current symptoms to reach a health diagnosis when a patient is ill. In order to recognize the combination of symptoms and history that points to a particular disease, the doctor's brain accesses memory of previous patients, as well as information that has been learned from books or other doctors. A neural network has the ability to mimic this type of decision-making process, and use a knowledge base of information, and a training set of practice cases, to learn to diagnose diseases.

2.6 Poisoning Attack

Systematic attack schemes for mounting poisoning attacks against machine learning algorithms used for medical datasets, and suggested countermeasures against them. A key feature of the proposed attack schemes is that they can be applied to a wide range of machine learning algorithms, even when the machine learning algorithm is unknown. Author evaluated the effectiveness of the attacks against six machine learning algorithms and five datasets (Thyroid Disease, Breast Cancer, Acute Inflammations, Echocardiogram, and Molecular Biology (Splice-junction Gene Sequences)), and ranked the algorithms based on their ability to withstand the attacks. They then presented countermeasures against these attacks and evaluated their effectiveness. Finally, we identified the machine learning algorithms that are easiest to defend [6].

3. Proposed Work

In healthcare, attackers may have varying motivations for poisoning training datasets, ranging from generally degrading the accuracy of the algorithm to biasing the results in a specific, targeted manner. As an example of targeted attacks, let us consider the Thyroid Disease dataset, in which data instances are associated with two classes: normal and hypothyroid. Targeted attacks might compromise the effectiveness of the machine learning algorithm either to prevent a hypothyroid diagnosis or to falsely lead to a hypothyroid diagnosis.

The application aims at providing medical assistance. The application will take as input text i.e. the symptoms and process it using the database. There are various parts of the project. The first part will be completing the basic requirements of the chatbot. It will take input and search in the database for the corresponding output and gives the result. For this search we will use various search-engines like special search engine, matrix engine which stores the basic conversation [3].

Figure 1: Disease Identification System

The second part of the application will involve helping the chatbot to learn. It means there maybe some symptoms which the chatbot doesn't know and if the patient mentions it then the chatbot stores it in the learning database. The chatbot also uses engines like previous conversation engine, auto-talk engine.

The third part of the application will involve applying poisoning attack on the dataset of chatbot and checking whether that dataset can be affected or not. Also the countermeasure algorithm will be applied to see till what extend the chatbot dataset can defend the applied poisoning attack. The proposed attack schemes can be used for targeted or non-targeted attacks. However, we describe the procedure in the context of targeted attacks [6].

4. Project Preparation

System to be developed is a chatbot that is capable of carrying on conversation with humans in natural language that is English. The main aim of the project is to create a system related to AI in medicines. Basically this application will be providing facility to the patients to have a conversation with a chatbot and ask him the diseases by providing symptoms to the bot. Many chat bots are available but they don't provide such facility for hospital applications. So we had thought of developing this application [1][2].

It is important to investigate whether machine learning algorithms used for healthcare applications are vulnerable to

security and privacy threats. The robustness of machine learning algorithms to noise in the training data has also been investigated to evaluate its effects on the decision making process. In healthcare applications, poisoning attacks are highly relevant because although manipulation of existing data in the training dataset may be difficult or impossible for attackers, addition of new data might be relatively easy. For instance, hindrance of a hypothyroid diagnosis may have life-threatening consequences due to delayed treatment. This may reduce trust in the machine learning algorithm.

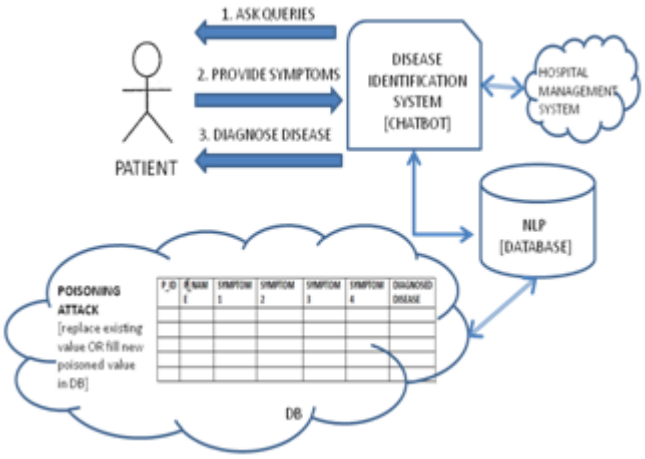

Figure 1: Disease Identification System with poisoning attack

On the other hand, a false positive classification may cause unnecessary concern. If poisoning attacks are detected, the user or owner of the dataset may take appropriate action, such as disregarding the results of machine learning or attempting to cleanse the dataset of the malicious data. From an attacker's perspective, it is therefore desirable to mount poisoning attacks such that they are difficult to detect. On the other hand, if such attacks are successful, the resulting erroneous conclusions may lead to serious adverse impact on people, institutions, and healthcare services.

To evaluate the proposed scheme, we will experiment with the chatbot dataset and show that the proposed attack is successful or not. Moreover, we show that one may obtain a surrogate dataset to mount the attacks, eliminating the need for access to the training dataset.

5. Attack Scheme

Our objective is to propose a generic and algorithm-independent attack scheme. In other words, the proposed attacks can be applied to a wide range of machine learning algorithms and medical datasets. In fact, the attacker does not even need to know the type of machine learning algorithm used to apply the proposed attack scheme. Furthermore, highly algorithm-specific attacks may be thwarted by simply changing the machine learning algorithm used. However, knowledge of the machine learning algorithm being used increases the efficacy of the attacks, as discussed later. In this attack model, we assume that the attackers have knowledge of the training dataset and use this knowledge to construct malicious data. In practice, this knowledge can be obtained either because the dataset is publicly available, or because the attackers have employed various means, such as eavesdropping on network traffic or compromising a system where the dataset is stored, in case security measures, such as the ones presented in, are compromised. However, the success of the proposed attacks is only dependent on the knowledge of the statistics of the training dataset [6].

In scenarios where gaining access to the training datasets is difficult, we present an alternative approach in which attackers construct a proxy training dataset drawn from the same distribution as the original dataset. This is possible since our proposed attacks are based on the statistics of the training dataset (and not the exact values of attributes within the dataset). By presenting artificial test instances as inputs to the targeted machine learning application and observing its responses, one can construct a "proxy" dataset that can be used to mount the attack [6]. Moreover, in many cases, launching poisoning attacks may be much easier than launching general causative attacks in which modifications to current instances are required.

References

[1] A Fast Multiple Pattern Matching Algorithm using Context Free Grammar and Tree Model, G.Phanindra, K.V.V.N. Ravi Shankar, P.DeepakSreenivas.
[2] Artificial intelligence in medical application: an exploration Wan Hussain Wan Ishak, FadzilahSiraj.
[3] One-sided algorithms for integrating empirical and explanation-based learning. Wendy E. Sarrett and Michael J. Pazzani.
[4] "Security evaluation of pattern classifiers under attack," B. Biggio, G. Fumera, and F. Roli, IEEE Trans. Know. Data Eng., pp. 984–996, Apr. 2014.
[5] "Poisoning attacks against support vector machines," B. Biggio, B. Nelson, and P. Laskov, in Proc. Int. Conf. Machine Learning, 2012, pp. 1807– 1814.
[6] Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare" Mehran Mozaffari-Kermani, Member, IEEE, Susmita Sur-Kolay, Senior Member, IEEE, Anand Raghunathan, Fellow, IEEE, and Niraj K. Jha, Fellow, IEEE.
[7] http://en.wikipedia.org/wiki/AIML
[8] http://www.cnlp.org/publications/03nlp.lis.encyclopedia.pdf
[9] http://www.informatik.uni-freiburg.de/~ki/papers/skocaj-etal-iros2011.pdf
[10] http://www.jcheudin.fr/pdf/2011_international_conference_agents_artificial_intelligence.pdf
[11] http://dl.acm.org/citation.cfm?id=1780909.1780989&coll=DL&dl=GUIDE

Author Profile

Avinash Shrivas completed BE in computer engineering in the year 1997 from Amravati university. He worked as lecturer in engineering colleges. He completed M. Tech in the year 2009 from NMIMS university. He is working as Assistant Professor in Vidyalankar Instituite of Technology, Mumbai. He is having a complete teaching experience of 18 years at UG and PG Level. His area of research is Artificial Intelligence.

Prasanna Shivaji Shinde completed BE in computer engineering in the year 2013 from Mumbai university. She is pursuing ME from Vidyalankar Institue of Technology.