Breast Cancer Diagnosis

Na Song

*School of Computing, Clemson University*

nsong@g.clemson.edu

***Abstract*—The rapid development of Internet of Things (IoT) has triggered more and more smart devices to be connected into the Internet. One typical smart device, smart home device, have been used by a large number of users. However, the wide deployment of these smart devices also raised security and privacy issues. In order to understand well the security vulnerabilities of IoT devices, in this group project, we apply a case study of smart plug system to analyze several attacks, such as device scanning attack, brute force attack, spoofing attack and firmware attack. Finally, we also present several defense strategies to mitigate the risks from the vulnerabilities of smart plug system.**

***Keywords***— **Internet of Things, IoT, Security, Vulnerabilities, Smart Plug**

1. Introduction

According to the Centers for Disease Control and Prevention(CDC), breast cancer is the most common type of cancer for woman regardless of race and ethnicity(CDC,2016). Around 220,000 women are diagnosed with breast cancer each year in the United States (CDC, 2016).

The diagnosis of breast tumors has traditionally been performed by a full biopsy an invasive surgical procedure, in the recent years, by using and extending the results from the fields of machine learning, statistics, image processing and optimization, highly accurate diagnosis of breast can be done even by untrained users. Using software system helps doctors to make accurate judgment is the tendency of the cancer diagnosis.

In this article, some machine learning algorithms will be applied to already processed images to get accurate judgment.

1. Dataset

The dataset is come from UCI Machine Learning Dataset, which called Breast Cancer Wisconsin (Diagnostic) Data Set. There are 569 instances in the dataset, and for each instances, there are 32 attributes to discribe it. All these features are computed from a digitized image of a fine needle aspirate(FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. There are both qualitative and quantitative features in the dataset. All feature values (for quantitative features) are recoded with four significant digits. As for the class distribution, there are 357 benign and 212 malignant.

Here is a brief introduction of each attribute.

* ID number

The ID of each sample.

* Diagnosis

M means malignant, B means benign. This attribute is the output.

The rest of the attribute are features, and they are all quantitative parameters. The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

* radius

mean of distances from center to points on the perimeter. radius\_mean, radius\_se, radius\_worst are the mean, the standard error and the worst of the radius.

* texture (standard deviation of gray-scale values)
* perimeter
* area
* smoothness (local variation in radius lengths)
* compactness
* concavity (severity of concave portions of the contour)
* concave points (number of concave portions of the contour)
* symmetry
* fractal dimension ("coastline approximation" - 1)

1. Exploratory Data

## *Overview Of Dataset.*

First of all, we count the diagnosis results as Fig.1.

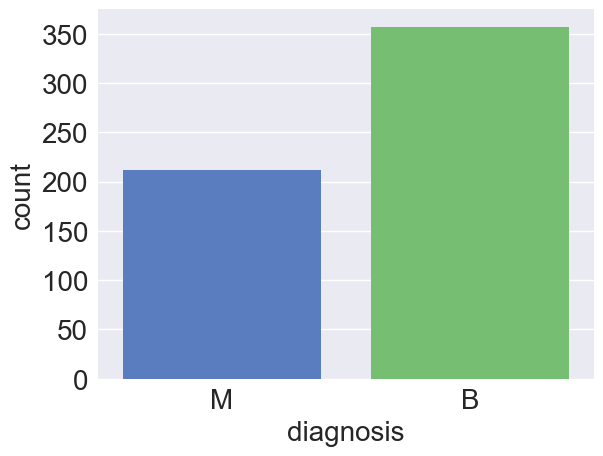


Fig.1 count for the diagnosis results.

There are 357 number of Benign and 212 number of Malignant.

In order to explore the influence of each feature on the output, at the first step, checking the difference in each feature’s value in different class.

Since the differences between values of features are very high to observe on plot, so before ploting, we need to normalizaion or standirdization. Here, we use the funcion below to normalize the values.

Then dividing the 30 features into 3 groups by mean, standard error and worst. And ploting the value and the classification of each group as Fig.2.

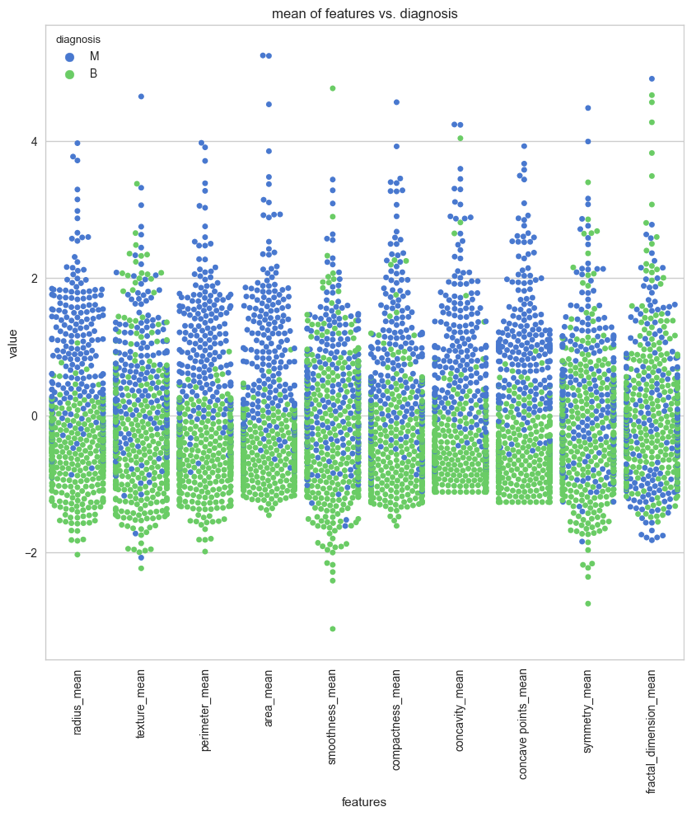


Fig.2 value distribution vs. the classification based on each feature.

From the plot above, we can find some useful information about features, for example, on the one hand, in perimeter\_mean feature, median of Maligant and Benign looks like seperated so it can be good for classification. However, in fractal\_dimension\_mean feature, mean of the Malignant and Benign does not looks like separated so it does not give good information for classificaion. On the other hand, we can see that the value distributions of radius\_mean and perimeter\_mean look similar with each other, which inspire us to explore whether they are correlated with each other or not.

And we can plot the same swarm plot for the other two groups: the standard error and the worst group.

## *Correlation Of Features*

As we can see in Fig.2 that there are some features which are similar with each other. In order to compare these feature deeper, we use Pearsonr value to access the correlation, where 1 is the highest. In Fig.3, we give an example.

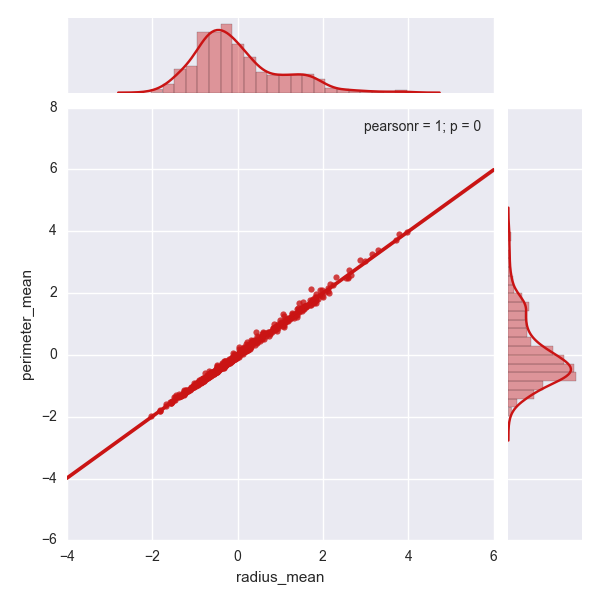


Fig.3 Pearsonr value of radius\_mean and perimeter\_mean.

The pearsonr value between radius\_mean and perimeter\_mean is 1, so they are high correlated.

One step futher, in order to see the observe all correlation between features, the heatmap of the Pearsonr value can be used to show the correlation between each features as Fig.4 shows.

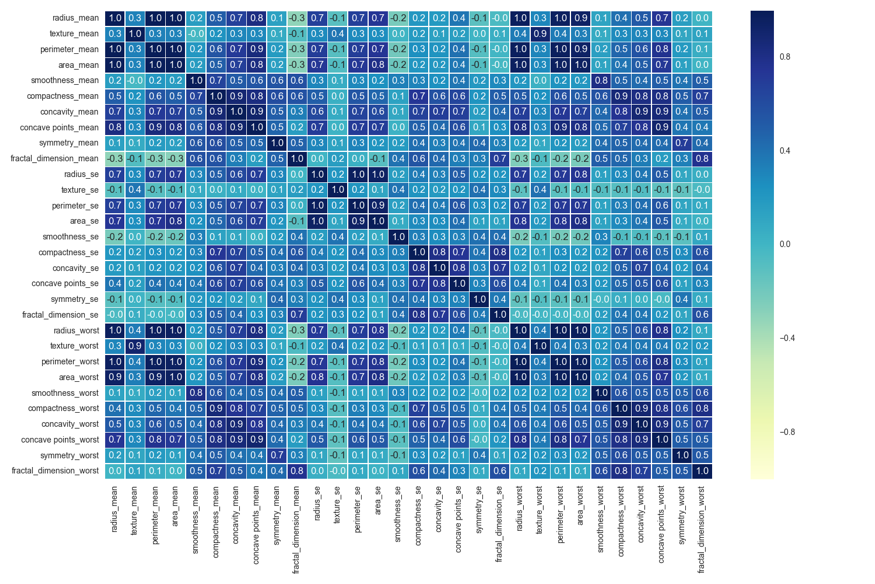


Fig.4 correlation heatmap for all features.

As it can be seen in heat map figure, **radius\_mean**, **perimeter\_mean** and **area\_mean** are correlated with each other so we will use only one of them. Here, summarizing the correlated features in the table1.

|  |  |  |
| --- | --- | --- |
| features | correlation | |
| radius\_mean, perimeter\_mean, area\_mean | | 1.0 |
| compactness\_mean, concavity\_mean  concavity\_mean, concave points\_mean  compactness\_mean, concave points\_mean | | 0.9  0.9  0.8 |
| radius\_se, perimeter\_se, area\_se | | 1.0 |
| radius\_worst, perimeter\_worst, area\_worst | | 1.0 |
| Compactness\_se, concavity\_se  concavity\_se, concave points\_se  compactness\_se, concave points\_se | | 0.8  0.8  0.7 |
| compactness\_worst, concavity\_worst  concavity\_worst, concave points\_worst  compactness\_worst, concave points\_worst | | 0.9  0.9  0.8 |
| Texture\_mean, texture\_worst | | 0.9 |
| Area\_worst, area\_mean | | 1.0 |

Table 1 high correlation value table

For the features that correlated with each other (the correlation value is 1.0), we can choose one of them to represent them, that means we can drop **radius\_mean, perimeter\_mean, radius\_se, perimeter\_se, radius\_worst, perimeter\_worst and area\_worst.**

1. Model selection and validation

## *Models with simply dropping features.*

In this part of models, we just drop the features that have high correlation values with each other. **Compactness\_mean, concavity\_mean** and **concave points\_mean** are correlated with each other. Therefore I only choose **concavity\_mean**. in **Compactness\_worst, concavity\_worst** and **concave points\_worst** so I use **concavity\_worst**. **Compactness\_se**, **concavity\_se** and **concave points\_se** so I use **concavity\_se**. **texture\_mean** and **texture\_worst** are correlated and I use **texture\_mea**n.

In order to quantify the relationship between the distribution of the features’ values with the diagnosis result(M/B), Fisher's score is used to show the distinction of different features. Since we have removed some of the features in the previous step, so that they will not be considered.

For a feature , the value of sample’s feature is . So that:

(1)

1. Conclusion

IoT system is similar to the traditional embedded system that it derives from. But IoT device is subject to be loaded with se

REFERENCES

1. Gartner, “8.4 billion connected ’things’ will be in use in 2017, up 31 percent from 2016,” http://www.gartner.com/newsroom/id/3598917, February 2017
2. D. Boswarthick, O. Elloumi, O. Hersent, M2M Communications: A Systems Approach, first ed., Wiley Publishing, 2012.
3. O. Hersent, D. Boswarthick, O. Elloumi, The Internet of Things: Key Applications and Protocols, second ed., Wiley Publishing, 2012.
4. Jacobsson, Andreas, Martin Boldt, and Bengt Carlsson. "A risk analysis of a smart home automation system." Future Generation Computer Systems 56 (2016): 719-733.
5. /DEV/TTYS0, “Hacking the D-link DSP-W215 smart plug,” http://www.devttys0.com/2014/05/hacking-the-d-link-dsp-w215- smart-plug/, 2014.
6. Lubomir Stroetmann, "Reverse Engineering the TP-Link HS110",https://www.softscheck.com/en/reverse-engineering-tp-link-hs110/, 2016 .
7. Wurm, Jacob, et al. "Security analysis on consumer and industrial iot devices." Design Automation Conference (ASP-DAC), 2016 21st Asia and South Pacific. IEEE, 2016.
8. Hernandez G, Arias O, Buentello D, Jin Y. Smart Nest Thermostat: A Smart Spy in Your Home. In: Proceedings of Black Hat 2014. Las Vegas. p. 1–8. 14
9. Arias, Orlando, et al. "Privacy and security in internet of things and wearable devices." IEEE Transactions on Multi-Scale Computing Systems 1.2 (2015): 99-109.
10. Rahman, Mahmudur, Bogdan Carbunar, and Madhusudan Banik. "Fit and vulnerable: Attacks and defenses for a health monitoring device." arXiv preprint arXiv:1304.5672 (2013).
11. Rouf, Ishtiaq, et al. "Neighborhood watch: security and privacy analysis of automatic meter reading systems." Proceedings of the 2012 ACM conference on Computer and communications security. ACM, 2012.
12. Zuo, Chaoshun, et al. "Automatic Forgery of Cryptographically Consistent Messages to Identify Security Vulnerabilities in Mobile Services." NDSS. 2016.
13. Lin, Jie, et al. "On false data injection attacks against distributed energy routing in smart grid." Proceedings of the 2012 IEEE/ACM Third International Conference on Cyber-Physical Systems. IEEE Computer Society, 2012.
14. Ronen, Eyal, and Adi Shamir. "Extended functionality attacks on iot devices: The case of smart lights." Security and Privacy (EuroS&P), 2016 IEEE European Symposium on. IEEE, 2016.
15. Pammu, Ali Akbar, et al. "Interceptive side channel attack on AES-128 wireless communications for IoT applications." Circuits and Systems (APCCAS), 2016 IEEE Asia Pacific Conference on. IEEE, 2016.
16. Park, Jungmin, and Akhilesh Tyagi. "Using Power Clues to Hack IoT Devices: The power side channel provides for instruction-level disassembly." IEEE Consumer Electronics Magazine 6.3 (2017): 92-102.
17. Song, Houbing, Glenn Fink, and Sabina Jeschke, eds. Security and Privacy in Cyber-Physical Systems: Foundations, Principles and Applications. John Wiley & Sons, 2017.
18. Koeberl, Patrick, et al. "TrustLite: A security architecture for tiny embedded devices." Proceedings of the Ninth European Conference on Computer Systems. ACM, 2014.
19. Eldefrawy, Karim, et al. "SMART: Secure and Minimal Architecture for (Establishing Dynamic) Root of Trust." NDSS. Vol. 12. 2012.
20. Ling, Zhen, et al. "Security Vulnerabilities of Internet of Things: A Case Study of the Smart Plug System." IEEE Internet of Things Journal (2017)