Application of Machine Learning Techniques to Situational Risk Assessment Based on Accident Database

Ryuta Watanabe, Keisuke Yamazaki, Tsuyoshi Nakajima
Department of Information Science and Engineering
Shibaura Institute of Technology, SIT
Tokyo, Japan

e-mails: {ma18109, ma17123, tsnaka}@shibaura-it.ac.jp

***Abstract*— It is challenging to prevent various kinds of offenses, such as bicycle theft or street snatching. Machine learning techniques, like Support Vector Machine (SVM) and Bayesian Network (BN), are considered to be promising technologies to assess the risk of such offenses. However, applying these technologies is not easy and some problems include preparation of missing data, providing reasons, and multi –level classification of risks. In this paper, we propose a method to solve these problems. We applied our proposed method on an example of risk information provision system on bicycle parking lots; the results showed the effectiveness of the proposed method.**

Keywords; Machine learning; Support Vector Machine; Bayesian Network; Risk Assessment.

I. INTRODUCTION

It is challenging to prevent various kinds of offenses, such as bicycle theft or street snatching. PredPol, which predicts crimes by using data of past offenses, has been used in the city of Santa Cruz, California, USA, and, as a result, contributed to decreasing the crime rate of that city [1].

The risk of accidents for a specified situation should be assessed using many factors, and so it is difficult to model crime occurrence mechanism mathematically. Machine learning techniques like Support Vector Machine (SVM) and Bayesian Network (BN) are considered to be promising technology for this purpose.

However, we have the following three problems for applying the techniques. The first problem is preparation of missing data. Machine learning techniques require as learning data both accident data that occurred in the past and non-accident data that did not occur. The problem is that, in general, non-accident data do not exist in the database, and, in addition, accident data in the database often have missing items in it. We should prepare such missing data and data items. The second problem is provision of reasons. Most machine learning techniques do not present reasons for the assessment results, which makes the user is doubtful of the results because he does not recognize why the situation is unsafe. The third problem is multi-level classification of risks. Machine learning techniques can classify the situation into safe or unsafe. Such a classification alone could not satisfy the user needs for determining what behavior to take. Although some techniques can provide the probability for the result, it is unlikely to be the real probability for the occurrence of the accidents due to the first problem.

We propose a method to solve the above three problems when using machine learning techniques to assess the risk of a specified situation based on accident database. As target machine learning techniques, we use SVM [2] and BN [3].

The rest of the paper is structured as follows. Section II describes an Internet of Things (IoT) system, as an example to which the proposed method is applied. Section III addresses the three problems, and the Section IV provides the method for solving them. We conclude the work in Section V.

II. A RISK INFORMATION PROVISION SYSTEM ON BICYCLE PARKING LOTS

This section describes an IoT system, as an example to which the proposed method is applied: Risk information provision system on bicycle parking lots [4], called RaBiPL.

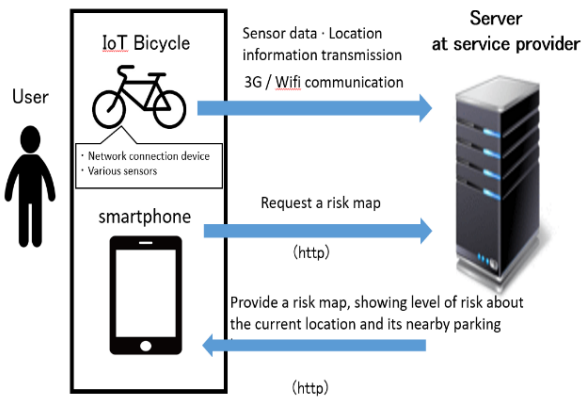

Figure 1. Risk information provision system on bicycle parking lots (RabiPL)

Figure 1 shows what devices the system uses and how it works. The system provides risk information for a specified situation to prevent from bicycle theft. This system works with IoT bicycles [5], which have various sensors and 3G/WiFi connection. The server of the system at the service provider site gets environmental information, based on which it assesses the risk of theft for the situation. The results are provided to the user's smart device upon request. The

server performs the risk assessment by analyzing the environmental information stored on the server.

TABLE 1. ENVIRONMENT INFORMATION

Information to be obtained		Method
1) Properties of bicycle		Obtaining from the information the user registered
2) Regional characteristics		Using publicly available database to extract information on specific region (using GPS sensor)
3) Time (hour, day, season)		Obtaining from the time the user request issues
4) Information on surrounding environment	Pedestrian traffic	Using human detection rate (using human sensor)
	Weather	Querying the weather service about weather information on specific region (using GPS sensor)
	Lightness	Obtaining directly from sensor data (using illuminance sensor)
	Temperature and humidity	Obtaining directly from sensor data (using temperature / humidity sensor)

Environmental information includes:
(1) Properties of bicycle, such as model and price,
(2) Regional characteristics, such as the number of crimes,
(3) Time, such as time zone and day of week, and
(4) Surrounding environment, such as pedestrian traffic.

TI shows the four types of environmental information and the method to obtain it.

III. RESEARCH PROBLEMS

A. *Problem1: Preparetion of missing data*

Machine learning techniques require as learning data both accident data that occurred in the past and non-accident data that did not occur.

The problem is that in general, non-accident data do not exist in the database, and in addition, accident data in the database often have missing items in the data.

In the case of RabiPL, their victims report many bicycle thefts, and their data are available on the Web [6], but each data may have missing elements, such as bicycle's color, and weather in which it occurred. Of course, there is no data for the situation where bicycle thefts did not occur.

B. *Problem2: Provision of reasons*

Most machine learning techniques do not present reasons for the assessment results, which makes the user doubtful of the results because he does not recognize why the situation is unsafe.

In the case of RabiPL, if the system informs its user that the risk that bicycle theft will occur in his situation is high, only the result may not convince him to stop parking there.

Problem3: Multi-Level Classification of risks

Machine learning techniques can classify the situation into safe or unsafe. Such a classification alone could not satisfy the user needs for determining his behavior to take.

Although some techniques provide probability of the result, it is not a true value for the accidents. Therefore, it is necessary to show how high the risk is at multiple levels so that the user can easily decide what to do.

In the case of RabiPL, users want to know that his situation is not risky, a little risky, risky, or very risky so that he can make a decision on whether he will park or not, considering his need to do it.

IV. PROPOSED METHOD AND ITS APPLICATION

A. *Preparation of missing data*

We propose a method to solve the three problems mentioned in Section 3 and describe its application.

1) Proposed method for preparation of missing data

To solve Problem 1, we propose the following method to prepare accident data with missing data items and non-accident data.

- Missing data items in accident data:
  - What can be obtained with sensors: field work
  - Others: randomly created value
- Non-accident data (the same number as that of the accident data) :
  - Place (Randomly selected from where the accidents did not occurred)
  - Time (Randomly selected)
  - Others:
    - What can be obtained with sensors: field work
    - Others: randomly created value

"Field work" is to go where the specified situation can be realized to collect data by using sensors.

2) Application to RabiPL

- Missing data items for theft occurrence data:
  - Using the theft data on the Web [6]
  - Obtaining regional characteristics from a Web service by searching by the current location
  - Missing data:
    - Information on surrounding environment (traffic lights, illuminance, etc.): obtaining the data items by "field work" at the place where theft occurred
- Non-theft occurrence environment data:
  - Randomly selecting the same number of non-theft data where no theft occurred in the publicly available list of the bicycle parking lots
  - Getting location of the lots
  - Randomly selecting time and properties of the bicycle
    - Obtaining regional characteristics from location with the Web service
    - Collecting non-theft environmental data with field work

In the "field work", we collected data for five minutes to use the average value of the measured sensor data.

Using the above method, we made 25 theft occurrence environmental data in the perfect form and

created 25 non-theft data, and we were able to prepare 50 cases in total.

B. *Provision of reasons*

To solve Problem 2, we propose the following method to provide reasons and describe its application.

1) *Proposed method for providing reasons*

We adopt Bayesian Network [3] as the applied machine learning technique, which is a probabilistic graphical model (a type of statistical model) that represents a set of variables and their conditional dependencies via a Directed Acyclic Graph (DAG). [6].

When using BN, intermediate nodes are chosen to explain the reasons for the assessment results, which have been an established common sense through statistical analysis conducted so far.

Then, after the assessment, choose the factors from those with higher risks than thresholds as the reason of the assessment.

2) *Application to RabiPL*

From the analysis on the theft of bicycles [7], we select the following four major factors as the intermediate nodes of the BN.

- Existence of those who want to steal
- Low possibility of detection
- Attractiveness of the bicycle
- Easiness to steal

Then, we link all the factors to the intermediate nodes. With regard to RabiPL, we linked all the environmental information items to the four intermediate nodes, as shown in Figure 2.

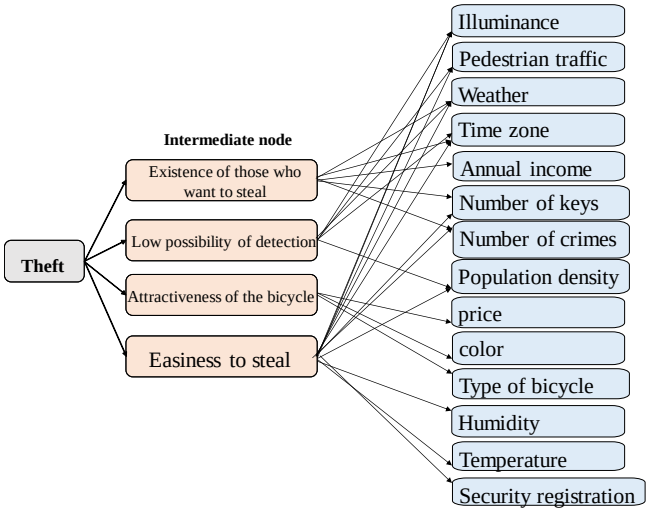

Figure 2.Relationship between intermediate nodes and environment information

Figure 3 shows an example to show the reason of the assessment result.

- The user wants to know why theft risk is high.
- The system selects the intermediate nodes that are the major contributors of the result. In this case, two nodes: "Low possibility of detection" and "Easiness to steal" are chosen.
- The system selects the environmental information items that contribute to the probability of the intermediate nodes and exceed the threshold. In Figure 4, two items: "Pedestrian traffic" and "Time zone" are chosen.

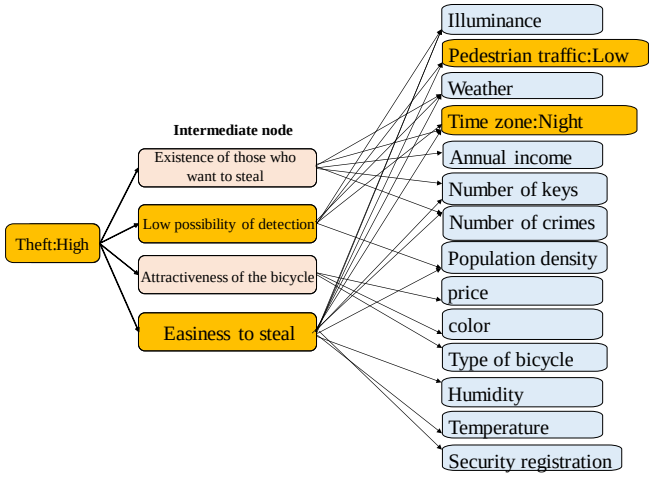

Figure 3.Relationship between intermediate nodes and environment information 2

- The system provides abstract and detailed reasons for the user, by using the intermediate nodes and environmental information items, as shown in Figure 4.

Figure 4. Evidence presented example

C. *Multi-Level classification of risks*

To solve Problem 3, we propose the following method to provide multi-level classification of risks and describe its application.

1) *Proposed method for multi level risk classification*

We also adopt Bayesian Network as the applied machine learning technique, which provides a probability value with the result. As mentioned before, the value is not the real probability of the occurrence of the accidents because of the Problem 1.

Therefore, we propose the following method to classify the result into N levels based on its probability value.

n: the number of data items.

d_k: k th data item $(k=1,..,n)$

a) Create a sequence S_d by sorting in descending order of probability values

$$S_d = (d_1, d_2, \cdots, d_n)$$

Where $P(d_i) < P(d_{i+1})$

b) Divide the sequence into N as follows.

$$S_i = (d_{l^i}, \cdots, d_{m^i})$$

Where $i = 1, \ldots, N$

$$l^i < \frac{n}{N}(i-1)$$

$$m^i \geq \frac{n}{N} i$$

c) Let a data d be at level K when satisfying the following equation.

$$P(d_{m^{k-1}}) < P(d) \leq P(d_{m^k})$$

Where $P(d_{m^0}) = 0$

$$P(d_{m^N}) = 100$$

2) Application to RabiPL

In the case of RabiPL, we set the number of classification level to three: high, medium and low. The number of training data used was 30, including 15 theft data and 15 non-theft data.

Figure 5 illustrates how the proposed method determines threshold values to classify a probability value into a certain level. In Figure 5, the horizontal axis is the probability value of the learning data: d_i, and the vertical axis is the sequence number i. The sequence numbers: 10 and 20 divide the set of learning data into three levels, and so the corresponding probability values P1 and P2 can be used as the thresholds that classify a probability value into a certain level: low, medium, or high.

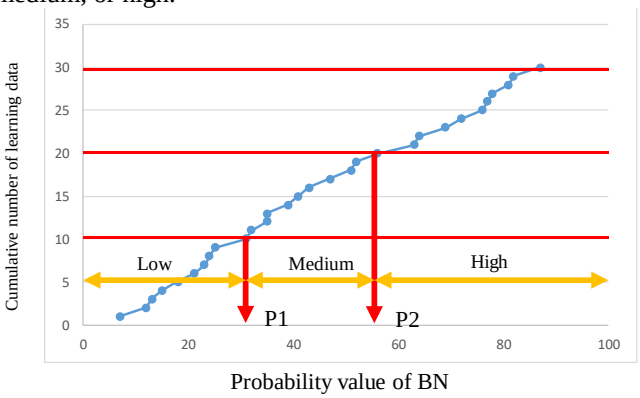

Figure 5. How to determine the risk

We use six theft data and four non-theft data as test data to evaluate the assessment results using the thresholds that the proposed method determined. Based on this data, the evaluation was made on multiple levels of risk determination.

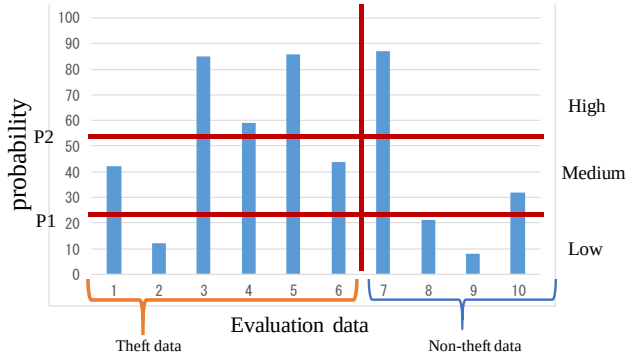

Figure 6. Evaluation on risk assessment

Figure 6 shows the probability values for test data, in which data 1-6 are theft data and data 7-10 are non-theft data. From this result, five out of six theft data are classified into high or medium level, and three out of four non-theft data are classified into low or medium level. This result shows this method gives us a good classifier of the inferred results.

V. CONCLUSION

We proposed a set of methods to solve three problems that machine learning techniques have when they are applied to the risk assessment based on accident database. We applied the proposed methods to risk assessment for bicycle theft, which shows a good performance of the methods, although a relatively small number of data are used for the experiment.

The proposed methods should be applied to risk assessment for other types of accidents so that it can prove useful more widely.

VI. REFERENCES

[1] E. Huet, "Server And Protect: Predictive Policing Firm PredPol Promises To Map Crime Before It Happens," Forbes, No. 2, 2015.

[2] C.C.Chung, and C.Lin, "LIBSVM: a library for support vector machines," ACM transactions on intelligent systems and technology, Vol. 2, No. 3, Article 27, 2011

[3] Bayesian Network http://nlp.dse.ibaraki.ac.jp/~shinnou/zemi2006/zemi06-bayesnet.html

[4] K.Yamazaki and T.Nakajima, “A risk information provision system on bicycle parking lots,” IEEE International Congress on Internet of Things (ICIOT), pp. 162-165, 2017.

[5] Kiefer, Chris, and Frauke Behrendt, "Smart e-bike monitoring system: real-time open source and open hardware GPS assistance and sensor data for electrically-assisted bicycles," IET Intelligent Transport Systems, Vol. 10, No. 2, pp. 79-88, 2016.

[6] CSI bicycle police 24 hour: bicycle theft map (in Japanese): http://www.cycle-search.info/csi/tabid/56/Default.aspx.

[7] Johnson, Shane D., A.Sidebottom, and A.Thorpe, “Bicycle theft,” Washington, DC: US Department of Justice, Office of Community Oriented Policing Services, 2008.