Data Mining in Detection of Customs Declaration Frauds

Zijie Fan

Seizure Rate

4.061%

under 5% inspection rate

Data Mining(DM)

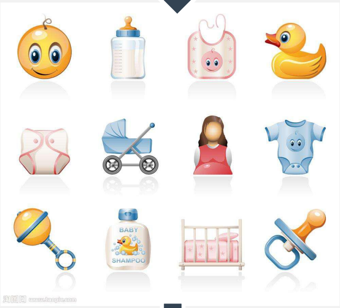
In a narrow sense, to use statistical and computer model to explore patterns concealed in data, thus to generate useful result(prediction).

content

Introduction Case

Online Shopping

recommendation based on what you have browsed



input data of browsed commodities in the past

train model to classify the commodities into a certain group

output the result to recommend commodities in that group

Process of Data Mining

What to eat? (target)

Ingredients in hand (data)

Cooking techniques (model)

How it tastes? (result)

How do Customs deal with the accumulated data?

Writing reports on a certain topic to indicate the trend of trade to inform Customs officers.

Hard to deal with a great volume and variety of data.

PB scale （about 1,000,000 GBs） & data from inside and outside

Use the computers to automatically analyze these data and generate useful results that may help Customs management.

Case Study

Problem: Customs declaration frauds
e.g. low declaration price; misclassification

Target: Using data of past declarations and their corresponding inspection results to train a classification model, thus to analyze the risks and predict whether a future given declaration should be inspected

DATA

23,330 pieces of declaration
and inspection data
China Customs
in several days of 2016

each piece of data has 16 fields: mode of transportation, consigner, consingee, gross-prize, unit-price, inspection result, etc.

15.7% pieces of data are labeled fraud

Support Vector Machine(SVM)

$$x = x_0 + \gamma \frac{w}{||w||}$$

$$\max \frac{1}{||w||} \quad s.t., y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$$

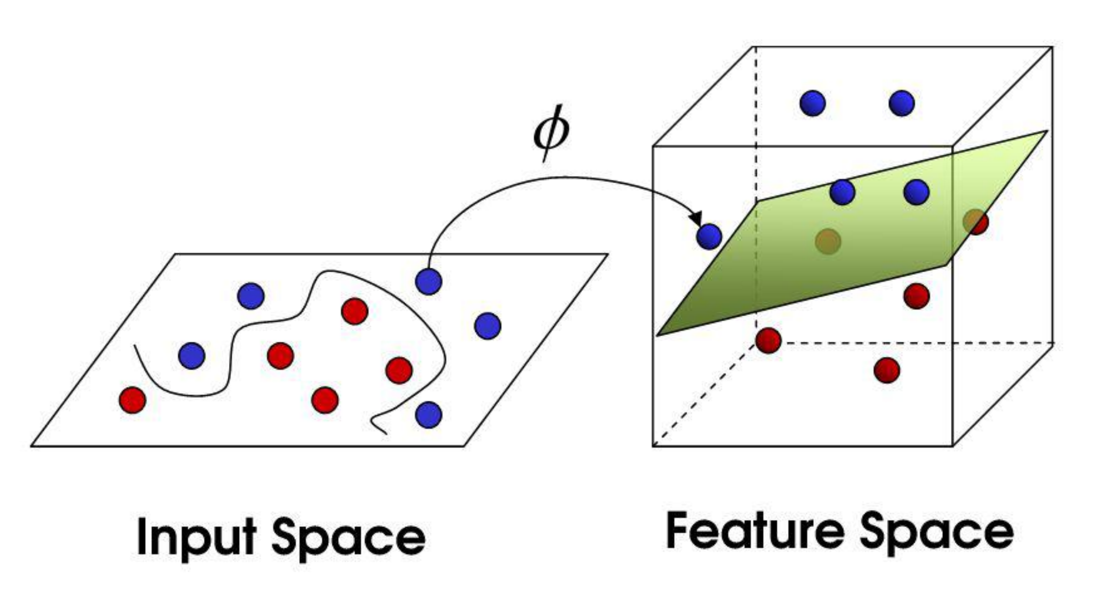
$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i \Big(y_i(w^T x_i + b) - 1 \Big)$$

$$\mathcal{L}(\mathrm{w}, \mathrm{b}, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(w^T x^{(i)} + b \right) - 1 \right]$$

$$\begin{aligned} \max_\alpha \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0, \end{aligned}$$

Support Vector Machine(SVM)

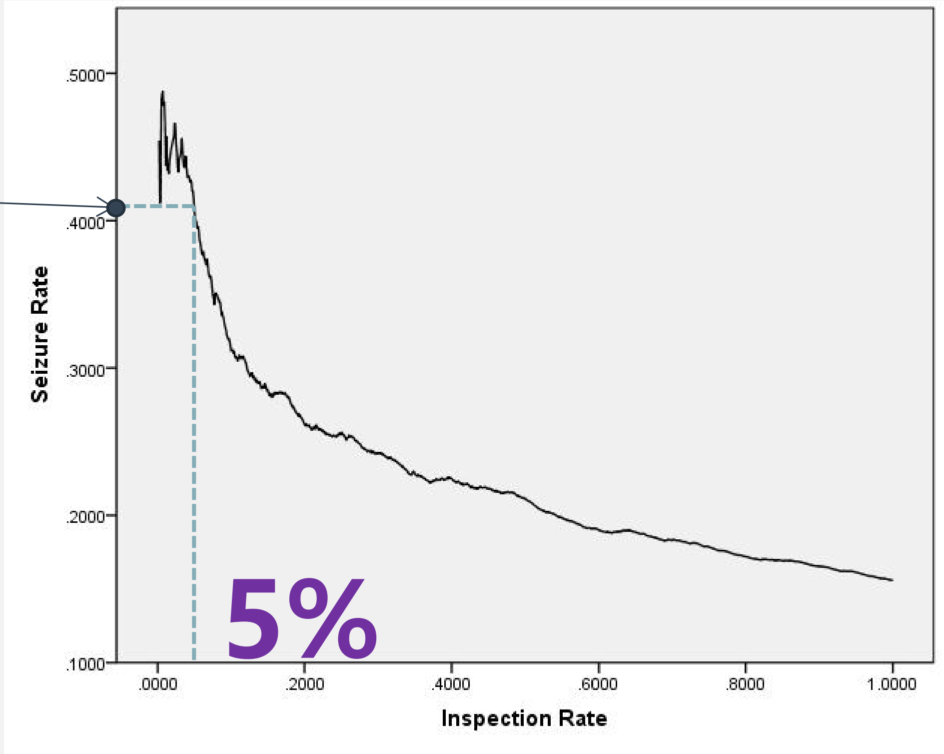
draw hyperplanes in high-dimensional feature space of data



Evaluation Index

Seizure-Inspection Curve

5% Seizure-Inspection Index



Data Mining Techniques

- divide data set into testing and training sets
- field transformation
- over-sampling
- regulate the parameters of the model

Result

Table Comparison of SVM with other common models(5% inspection rate)

Model	Seizure-Inspection Index	Training Time(second)
SVM	**40.61%**	152.62
CHAID	33.54%	3.59
Logistic Regression	34.11%	15.43
Bayesian Network	30.89%	1.17

40.61%
seizure rate!

*only declaration data are used to train this model: no data of intelligence and data from outside Customs(e.g. manifest) are used

Conclusions

Pros:

- Highly *effective* model in detecting Customs declaration fraud: 40.61% seizure rate under 5% inspection rate
- Highly *efficient* model: automatically analyze all declaration data of several days in about 2.5 minutes
- Advantageous method for Customs to use with the *gigantic data* accumulated by digital system
- Promising in *many fields* of Customs: Customs audition, passenger inspection, etc.

Conclusions

Cons:

- *Black Box*: the pattern of frauds detected by the model can hardly be interpreted by human
- No forthcoming *expert* in Customs: hire expert outside Customs or train Customs officers doing statistics jobs
- Investment in *hardware* for practical deployment: pioneer research still can be done

Suggestion

Researcher:

- Use *time sequence model* to reveal the trend of Customs declaration fraud

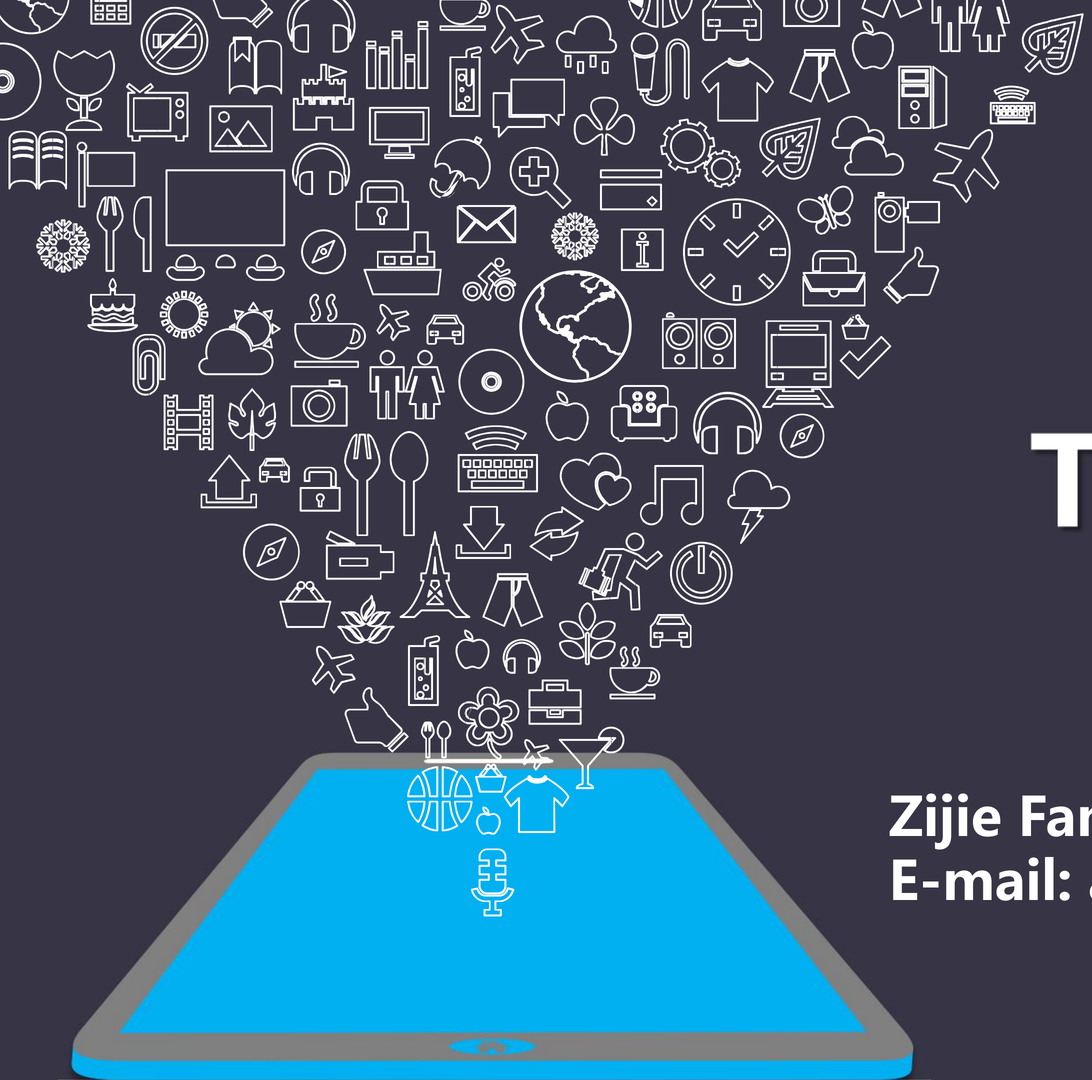
- Use data from *other sources*(e.g. manifest) to build a group of models

Suggestion

Administrator:

- lunch *programs* to find the possibility of combining Data Mining and Customs
- Cooperate with *IT business company* to find the possibility of practical deployment

Customs could take its advantage of high digitalization to pioneer the combining of Data Mining and public administration



Thank You

Zijie Fan, Shanghai Customs College
E-mail: awexv1@163.com