Introduction for Data Wrangling by Integrated Data Analysis Environment Padoc

*[1]Intelligent Mechanical Interaction System, University of Tsukuba

When analyzing business problems from data such as pricing, fraud detection, customer rating, etc., there is rare that appropriate analysis data is available. In these case analysts must extract data from sales or accounting systems which are indispensable in business. Generally, the data of the accounting system is stored at the detail level in the relational DB. The process of extracting data which implies the sign of the problem event from detail level data and formating data into the required format for the analysis tool is called preprocessing. This process is said that it is shared about 70 % of the entire process. On the other hand, the efficiency of the preprocessing has not progressed and this is dependent on the ability of the analyst. We think the reason why preprocessing is hard because preprocessing requires all of business knowledge, data shaping knowledge, and statistical knowledge. Since the support of the business side with business knowledge is essential for analysis from the detail level, the side responsible for data shaping and statistical knowledge needs to clearly present the data analysis process to the business side. We have developed Padoc(Platform of Analysis for Document), which is an integrated analysis environment for data analysis. We shows that Padoc is an effective tool when analyzing in cooperation with business side through comparison of various analysis tools.

1. Introduction

In our experience there are mainly following troublesome cases when we try analyzing business problem.

1. Awareness of the problem, but no data for the problem.
2. Holds only the latest data (transaction data)
3. Few identification of the problem event in the data

In the case of 1, data will be searched from the detail level of the accounting system , which are indispensable for business. Since the accounting system generally stores data in the relational DB to avoid duplication of data, for the extracting data properly a support of business side is absolutely necessary to recognize data. In the case 2, the business side can grasp the current financial situation by only transaction data, so do not pay attention to past data. However, historical data is needed to estimate the occurrence of problems from the past signs. In this case, it is necessary to introduce external data or wait until the data is accumulated. In the case 3, in general, the identification label is rarely completely satisfied, and the analytical model is divided into supervised and non-supervised analytical methods depending on the degree of satisfaction. The work process of selecting, integrating and shaping data from the detail level is called preprocessing, and it is often reported that this process is shared 60% to 80% of the entire process [New.York.Times 14] [Fuche 16] [Hameed 20] [Patil 18]. There are many tools[Aattenbury 17] to improve preprocessing efficiency, but users of these tools are expected analysts who are familiar with the business and still need on a lot of manual work [Hameed 20]. We think the reason why the preprocessing is hard because all of the following knowledge is required in the preprocessing.

- Business knowledge is required to understand detail level data
- Data shaping knowledge is required to integrate detail level data and shape data suitably for analysis tools
- Statistical knowledge is required to make a significant interpretation from the data

There is rare to have the above three knowledge, and if any of the knowledge is poor, many trial are spent until a significant result is obtained. Since the support of the business side is absolutely necessary to extract detail level data from the DB, we think that it is realistic and rational to collaborate with a so-called data scientists who have data shaping skill and statistical knowledge. For this cooperation data scientist side should use an easy-to-understand analysis tool that can share the recognition with the business side. We have developed the data analysis integrated development environment Padoc (Platform of Analysis for Document). We believe that this tool can present the data analysis processing to the business side in an easy-to-understand manner. However, in analysis that relies on the knowledge of the business side, there is a risk of missing data that the business side does not recognize. As a countermeasure in this case, there is a way of actively using the data search tool EDA (Explore Data Analysis) [Elansary 21]. But there is a more effective method that can detects the missing data by comparing with the model made by pattern recognition without relying on business knowledge. This method is described later in Section 7. The rest of the paper is as follows. In Section 2, we describe the problems of data preprocessing in recent years, In Section 3, we show that preprocessing requires 3 knowledge of business knowledge, data shaping skill and statistical knowledge, and In Section 4, we compare preprocessing tools, In Section 5, we show that Padoc is appropriate for analyzing business problems in coopera-

tion with the business side, and In Section 6, we describe the specific performance of Padoc.

2. Resent data preprocessing issue

In recent years, due to data sufficiency and computer performance improvement, machine learning models using Bayesian statistics and deep learning have shown remarkable results. But new problems are emerging. First, Python is too advanced that making the preprocessing by only Python is hard because this process is mainly shaping data depending on the business Knowledge. Second, a high-performance machine learning model can improve the apparent accuracy by learning even inappropriate data. But machine learning model which learn inappropriate data makes often accuracy deterioration in the business that use this model. Therefore, the importance of preprocessing to improve data quality is increasing more than before.

3. Rescent category of preprocessing

In recent years, the preprocessing is roughly divided into data preparation and data wrangling, and each has the following processes [Gill 17].

1. Data preparation
   (a) selecting and integrating data that are assumed to indicate a sign of the problem event from large amount of distributed detail level data in the DB
   (b) Interpretation of numeric code and category code
   (c) Correction or elimination contradictory data
   (d) Elimination of duplicate data that are made by data integration
   (e) Detection of the causes of lack data and elmination of lack data
   If the reason for the lack data is clear, lack data can be eliminated by dividing the data for that reason. If left unattended, accuracy will be deteriorate by change of satisfaction in business.
2. Data wrangling
   (a) Analysis of the relevance between the integrated data and the problem event (cross-validation or non-independence test)
   (b) Validation of a significant number of data to capture the problem event
   (c) Validation the independence of each record and column
   Independence is promised in general machine learning, and if independence is impaired, accuracy becomes deteriorate.
   (d) Eliminate inappropriate data (Avoid Leakage)
      i. Detected anomary data
      ii. Data generated by the problem event
      If this data is included, the apparent accuracy will be significantly improved.
      iii. Count up number by each time assigned
      Larger counts up data indicates the later data

Data preparation mainly requires business knowledge, and data wangling mainly requires statistical knowledge. The serious problem of preprocessing of business data is that data shaping skill, business knowledge and statistical knowledge are required at the same time, therefore the number of people who have these three knowledges is quite rare and this is mainly reason for not possible results. Regarding analysis tools, SQL that is good at only data shaping and Python that is good at only analysis, both are not enough. Required tool will work in simple and transparent processing for business side.

4. Comparison of well-known data preprocessing tools

Recent data preprocessing tools are divided into GUI type and program type[Molder 19]. The GUI type is used for only preprocessing with advanced wrangling tools, but the program type is used for both preprocessing and analysis by programing.

Table 1: Wrangling and Analysis Tools

Tool	Free	GUI	Wrangler	Analysis	IDE
SAS				○	○
Python	○		○	○	○
SQL	○				
Excel					
BigQuery				○	
Trifacta		○	○		
Talend		○	○		
Padoc	○			○	○

GUI:Graphic User Interface
Wrangler: Data Wrangling Tool
IDE:Integrate Development Enviroment

As mentioned above, the preprocessing of business data requires all business knowledge, data shaping skill and statistical knowledge, but we think that these existing data preprocessing tools focus on either knowledge as following.

1. Not free Trifacta can use various data formats such as Jison and Hadoop, and preprocessing scenario settings and advanced wrangling tools can be applied using the GUI. This tool seems that operations are easy for the business side to use because of visual operation. However, in actual work, visual operation works limited and requires advanced experts and manual work[Hameed 20][Molder 19]. This tool is suitable for data scientist side rather than the business side.

2. Business use Talend is kind of ETL (Extract Transform Load)tool. This tool has various GUI function for data extraction and data shaping. But the users of this tool are the system side rather than analysis side[Hameed 20].

3. Free Python provides various tools and advanced analysis models, but these are provided in an object-oriented method with hidden contents, the processing is not transparent, and Python requires too high technical skills for the business side. Python also has Pandas as a data editing library, but it provides only standard functions and is not suitable for large-scale data preprocessing. However, in recent years, advanced and various wrangling tools in Python has published such as missing value compensation and abnormal value detection as EDA (Exploratory Data Analysis)[Elansary 21], but has not been generally known.

4. Not free SAS provides a simple and simple-to-easy data editing environment like Padoc, so this tool can be used by business side who are good at programing, but SAS is too expensive and does not provide advanced analysis models. Recently, SAS can use embed simple Python code.

5. Excel is effective by tabulating within the only visible range of data, and this tool is suitable for business side which has low size of data. But it cannot perform advanced analysis such as machine learning.

6. Free SQL is convenient for data integration and editing. This SQL is suitable for business size because of ease to lean editing data.But it does not have an analysis function.

7. Not free BigQuery makes possible to edit data and analyze by coupling SQL and Python in the same environment because Python is too advanced for editing data. However, BigQuery is not easy to develop an analysis program because BigQuery is mainly used in Cloud and IDE is not equipped there.

If you want to introduce a free tool for data analysis, data integration and shaping can be used by free SQL and analysis can be used by Python. But data shaping and analysis are operated alternately, this switching operations increasingly become complicated. Google provided BigQuery as a solution to this complicity, but BigQuary is expense and required advanced technique in Python.

On the premise that there is a proper support from the business side, we believe that analysis based on their assumption about the business problem is more effective than based on exploration about business problem without their support. Therefore we provide Padoc as a simple data editing and standard analysis environment to analyze in cooperation with business side and data scientist side.

4.1 The premise to propose

The broad definition of data analysis is to extract knowledge from data, but the definition of the data analysis in this paper is making a model that predicts the occurrence of business problem events from data by machine learning etc., and estimation of the effect of the predictive model in the business. In addition, on the premise of data acquisition, there is past data and certain identification sign data about problem events, and proper data can be selected from the accounting system with the support of the business side. On this premise that there is a proper support from the business side, the procedure of selecting data which implies sign about business problem, and shaping data into a ordered format for analysis model are more rational and effective for both data scientist side and business side rather than working each other. Padoc provides the following performance and we believe that Padoc will work well in this cooperation.

- Alternating between data shaping and estimating analyzed results can be operated by Padoc integrated data analysis environment, and the shaping process is easy for the business side to understand.

- Data can be shaped by script of structured language in like C language style, and this description is easy to understand for the business side.

- Since data is limited to table type, the shaping process is clear and easy to understand for the business side.

- Padoc has the function to overlook about the problem event.

- In addition to standard analysis models, Padoc provides the external connection function with programs made based on business-specific logic.

5. Integrated data analysis Environment Padoc

In the following section the function of Padoc is shown by actual examples.

5.1 Overview of integrated data analysis environment

Similar to the integrated development environment of general programs, Padoc uses a platform as shown in the figure 1 (1) Main control panel for project management, etc. (2) Edit frame on which structured language is described for data shaping and analysis. (3) Log Frame on which ,if there is an error, a message is displayed (4) Data table Frame on which current data is showed (5) Plot Frame is equipped with an analyzed plot diagram. Padoc is a tool that interactively improves the accuracy of analysis referring to these Frames.

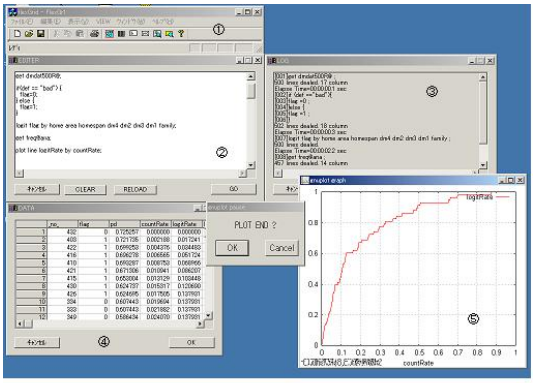

Figure 1: Overview of Padoc IDE (1)Main Control (2)Edit Frame (3)Log Frame (4)Current Data Table for Edit (5)Plot of Result

5.2 Data shaping

In Padoc, all data to be spaped is limited to table type data, and when you read the data from csv file or text file, you can refer the imported data as shown in the left table of the figure 3. On the edit Frame, you can shape the imported table-type data into the format required by the analysis tool. In the example below, in order to classify a donut-shaped zone on a two-dimensional plane by SVM [Plat 99], SVM requires proper points of the sign as +1 or -1 to learn division. The settings are made in the blue frame in the figure 2, and SVM is executed in the red frame. All variables that appear in the program are table columns names, and newly set variables are added to the table columns as shown in the figure 3.

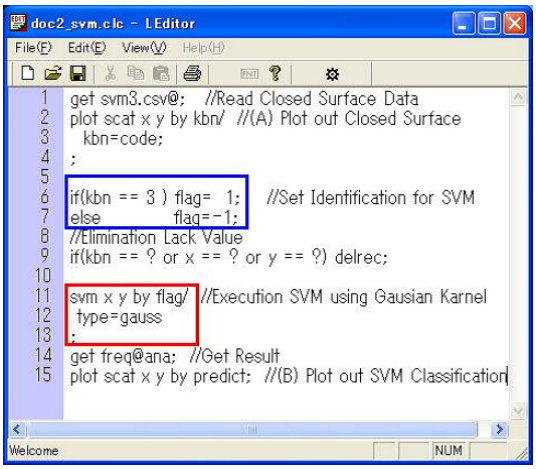

Figure 2: View of SVM script

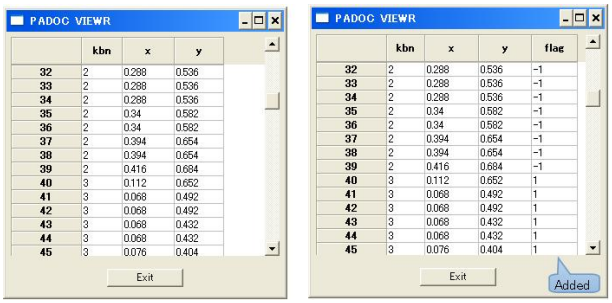

Figure 3: Left:Orignal Table Right:Added Column as flag

Figure 4 is the result of SVM prediction about donut zone indecated by binary values (+1, -1).

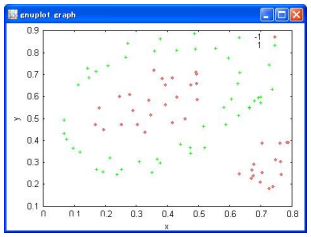

Figure 4: Classification by SVM for Donut Zone

Various functions are needed to shape data well. Padoc has more than 100 functions. The following are the main functions.

Table 2: Main Functions for Data Edit

Function	Content
merge	Merge two tables with a key column match
mxmult	Multiplication between tables
outrec	Condition select records from table
delrec	Condition delete records from table
unique	Eliminate duplicates for key column
julian	Translate date to total days
strsel	Extracting characters from a string
transpose	Transpose row and column of table

See below web cite for details
`http://www.padoc.info/sub/command_e.htm`

5.3 Adoption of Structured language in C style

Preprocessing needs combination of various functions to integrate distributed data and wrangle the data which includes leakage data. For such complicated processing, a structured language that well describes conditional sentences and repetitive sentences is adequate for preprocessing. Padoc can use a structured language in C style as shown in the figure 5, which makes it possible to express concisely from top to bottom even if there are conditional sentences and iterative processing, and this description is easy for the business side.

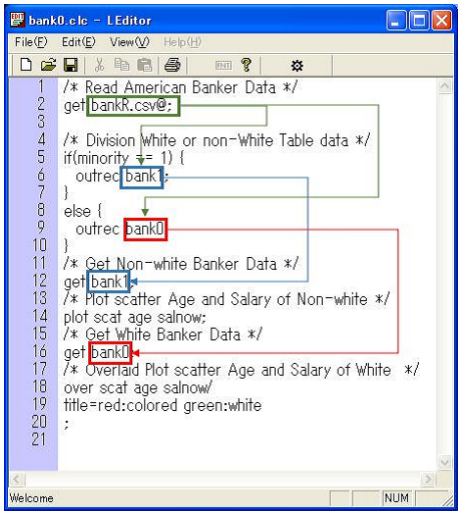

Figure 5: View of Edit Frame

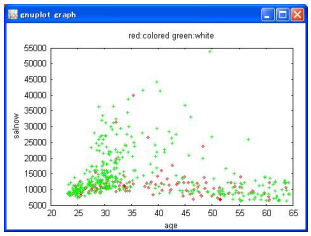

Figure 6: View of Plot Figure(Overlaped scatter plots show that young whites(green) are paid higher than Colored(red))

5.4 Data format is limited to table type data

In Padoc, all data to be shaped are limited to table type data only. As shown in the frame in the figure 5, table type data is the target of the structured language. By using the table type data, the transition process of the table type data can be clarified from top to bottom as indicated by the arrows in the figure 5. As an example, the script in the figure 5 analyzes whether white and non-white salaries of US bankers are unfair. The script separate white and non-white table data and scatter plots with salary and age for each. The superposition diagram of the figure 6 shows many scatter plots with high salaries of young white.

5.5 Interactive data analysis

Data shaping and analysis need to be able to be corrected repeatedly until making desired result, and correction, execution, and result reference need to be performed interactively. In Padoc, only the masked script lines as shown in the figure 7 can be executed by pressing the gear button at the top of the Frame. If execution makes an error,Log Frame instructs the error by messages as shown in the figure 8.

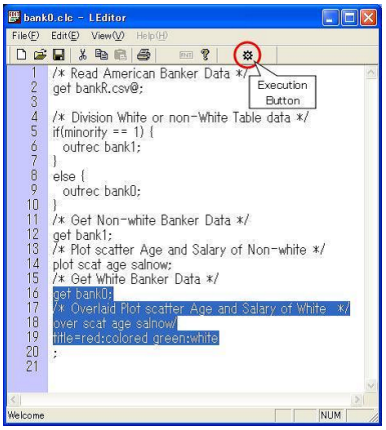

Figure 7: Execution only masked script lines

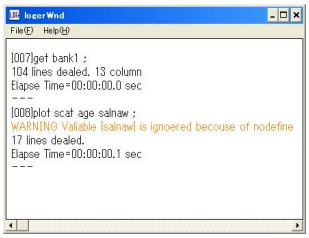

Figure 8: View of Log Frame

5.6 Overlook of data

As preprocessing, EDA (Exploratory Data Analysis) [Elansary 21] such as Trifacta [Aattenbury 17] and Python is very convenient because they can visually show the state of distribution, lack values, outliers, etc. by various operations or scripts. However, when the problem event is obvious, it is important to see that item of table data is how closely related to the problem event. Padoc has a function to display the increase / decrease of data values according to the occurrence rate of problem events in descending order, and this function is effective to search for the tendency of problem events. As an example, in the analysis of housing price estimation about figure 9 data which includes features of house, house price and higher-priced flag as items of the table data.The figure 10 shows the ranking item of table data in which high-priced houses exits.

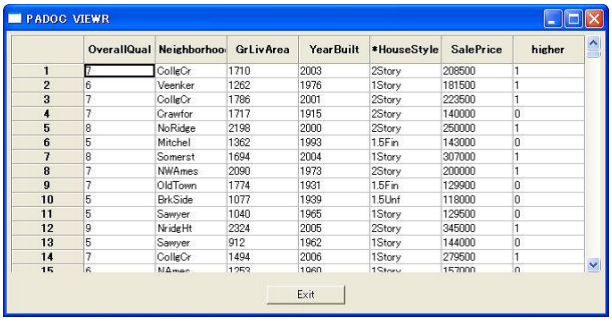

Figure 9: Example of House Price data

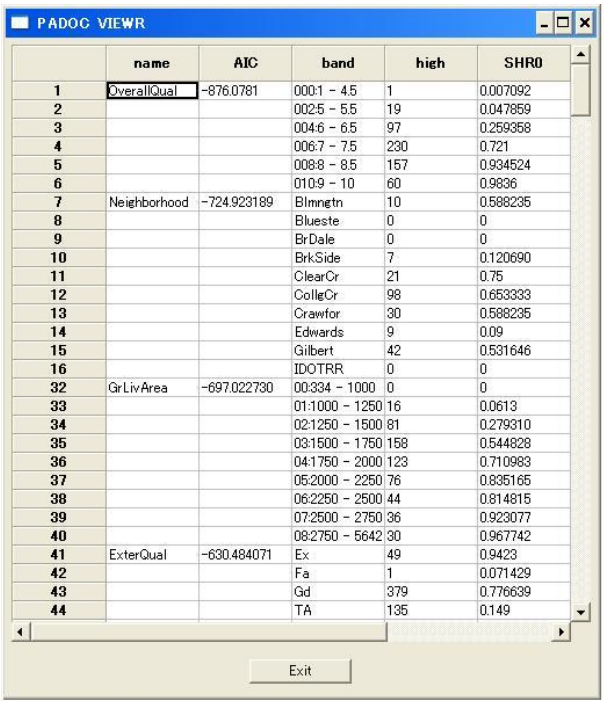

Figure 10: Ranking view for the target variable

If the item has numerical value, this function shows the relationship between the numerical value and the occurrence rate of the problem event. In figure 10 the first item 'OverallQuall' which is overall evaluation score shows that the composition ratio of high-priced is as higher in the SHR0 column as the more score value in the band column. The second item, 'Neighborhood', which is neighborhood towns shows that home prices depend not only on quality but also on the environment. This function makes you possible to see the relationship between the ranking item and the problem event as a overlook.

5.7 Providing analysis model

What is required of an integrated data analysis environment is the easiness of data shaping and various analysis models. The current Padoc has wide range of analysis models such as regression analysis, discriminant analysis, natural language analysis, time series analysis and planning method etc. Padoc provides analysis models mainly built by standard machine learning algorithm [Bishop 06] as shown in the table 3.

Table 3: Analysis Models of Padoc

Actual	Analysis	method1	method2	method3
○	Regression	linear	logit	
○	Classification	tree	svm	softmax
	Time series	arima	kalman	hmm
○	Survival analysis	cox	regtree	
	Natural Language	morpheme		
○	Optimize Plan	linear	quadoratic	
	Factor Analysis	prin	colabo filter	factor
	Graphical Model	ggm	Dijkstra	

Actual : Actual used in business

5.8 Interface of external defined analysis model

In a integrated data analysis environment, it is desirable to provide all analysis models, but depending on the problem, a business-specific solution may be more effective than the standard analysis model. Therefore, Padoc has a function to incorporate the external logic specialized for problems.

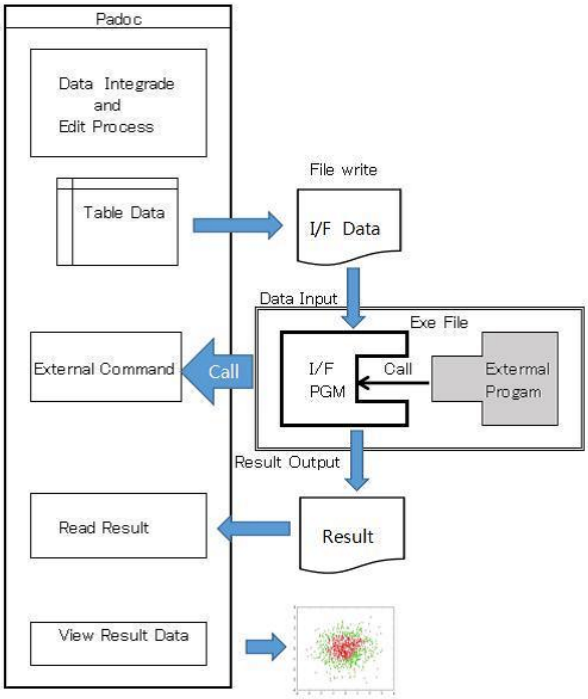

Figure 11: Relation of Padoc and External logic

As shown in the figure 11, at first the user makes an external logic (Specific Program) that needs to match the fixed I/F of Padoc. In cooperation, Padoc edits the data in a required format by the external logic, and calls the external logic by the command , and displays the result. As an example, we tried to incorporate the model of reinforcement learning which can path the simple maze [Sutton 14]. Figure 12 is this script. In this script the wall data is loaded, and the external logic of reinforcement learning is executed

in red frame, which instructs start and end position. To display result the wall positon is ploted in the blue frame and the route is ploted by superimposed view in the green frame. Reinforcement learning starts from the upper right and by maximizing the rewards it goes to the lower left. Figure 13 shows the result. The green dots indicate the walls and the red lines indicate the paths obtained.

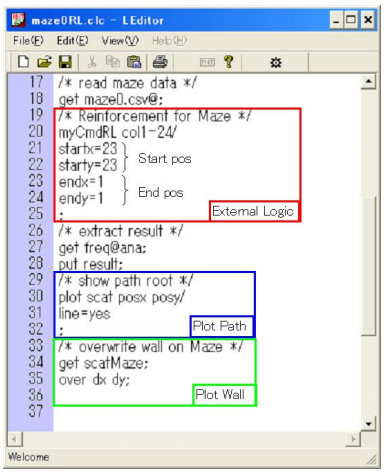

Figure 12: Script of imported RL module

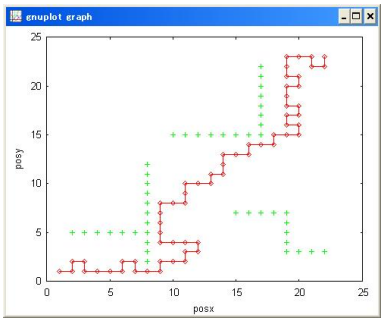

Figure 13: View of Result of RL

6. Adaption of Pattern recognition model

This paper proposes Padoc for the predict model about problem events on the promise of the business side support. However, there is a risk of miss signs about problem events that is not recognized by the business side. In this regard, there is a way of actively use of EDA tools, but we believe that comparison with a prediction model based on pattern recognition is effective. Even if the measure of accountability has improved[Scott 19] recently ,a patter recognition model is difficult to explain the prediction results such as Xgboost [Chen 16], but since the results can be obtained even with data including abnormal values or lack value due to insufficient preprocessing, we might detect significant data because business side thought this data was unnecessary for problem events. That is,as shown in figure14, if there is a data segment in which the occurrence rate is high when the pattern recognition model is used but the occurrence rate is low in the prediction model by the business side, there is a possibility that new knowledge can be obtained by examining the state of this data. In our experience, we got a lot of suggestions and reviewed our business knowledge.

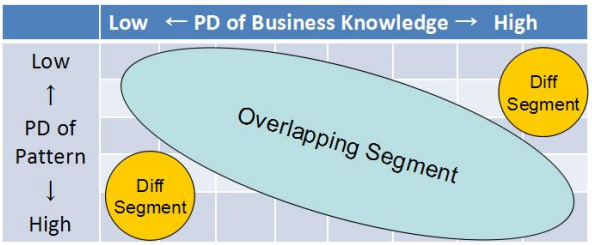

Figure 14: Cross Probability Distribution of models based on Business Knowledge and Pattern Recognition

7. Conclusion

As a premise of this paper, the scope of analysis is from making the prediction model for business problem events to the evaluation at business operation, and the data has the label of the problem event, and corporation with the business side to search data from the accounting system. This premise requires transparent and easy-to-understand work that can be easily understood and confirmed by the business side. We believe Padoc can respond in this regard. However, data analysis has a wide range of models for knowledge discovery such as testing, optimal planning, causal inference and prediction in various fields. Generally, the appropriate preprocessing tools and analysis models differ depending on the application field. EDA (Explorarity Data Analysis) is especially useful when the problem is vague and the signs are examined from the data. There are many Wrangling tools [Hameed 20] in the GUI type, and Python's rich wranling tools citeelan. If these will be refined and spread, we think that the efficiency of the preprocessing process will be improved.

References

[New.York.Times 14] New Yrk Times,`http://nyti.ms/1Aqif2X`

[Fuche 16] Fuche T., et al.,Data Wrangling for Big Data:Challenges and Oppotunities, In EDBT,2016

[Gill 17] Gill N. S.,Data Preprocessing and Data Wrangling in Machine Learning and Deep Learning,VIBLO Learning 2017

[VADA 17] Konstaninou N., et al.,The VADA Architecture for Cost-Effective Data Wrangling,dl.acm.org,2017

[Hameed 20] Hameed M., et al.,Data Preparation:A survety of Commercial Tools,SGMOD record(vol49.No.3),2020

[Patil 18] Patil M. M., et al.,A systematic Study of Data Wrangling,MECS,2018

[Aattenbury 17] Rattenbury T., et al.,Principles of Data Wrangling,O.REILLY,2017

[Molder 19] Molder H.,et al.,A Component-Based Approach to Traffic Data Wrangling,2019

[Elansary 21] Elansary M.,Data wragling & preparation automation,KU LEUVEN,2021

[Kazil 17] Kazil J., et al.,Data Wrangling with Python,O'REILLY,2017

[Bishop 06] Bishop,C.M Pattern Recognition and Machine Learning 7.1.3,Springer, 2006

[Plat 99] Platt, J. C., Fast Training of Support Vector Machines using Sequential Minimal Optimization,1999

[Sutton 14] Sutton R. et al.,Reinforce Learning:An Introduction,The MIT Press,2014

[Scott 19] Scott M, et al.,Consistent Individualized Feature Attribution for Tree Ensembles,arXiv:1802.03888v3,2019

[Chen 16] Chen, T., Guestrin, C.: XGBoost :A Scalable Tree Boosting System. arXiv2016.02754(2016)