Survey of Data Mining for Mechatronic Systems

Diploma Thesis

Xu Tian

Supervisors

O.Univ.-Prof. Dipl.-Ing. Dr.techn. Paul O'Leary

Ass.Prof. Dipl.-Ing. Dr.mont. Gerhard Rath

Montanuniversität Leoben

Chair of Automation

December 2014

Abstract

Data mining is a process of using various algorithms to transform an original data set, which may be affected by noise and missing values, into a form that can be analysed easier by human in order to extract information from it. This thesis gives an overview of the process and a brief introduction to commonly used algorithms. Among them symbolisation methods have some advantage for data mining. They allow convenient visualisation for human or automated search with symbolic queries, for example for repetitive pattern identification and discord detection. Especially the Symbolic Aggregate Approximation method allows efficient reduction of dimensionality and indexing with a positive semi-definite distance measure. After giving an overview, the thesis focuses on mining a real data set that was recorded on a production machine. Twenty sensors delivered values over more than a year resulting in a huge amount of approximately one billion measurements. For two exemplary sensors, the application of several algorithms is demonstrated, such as preprocessing, k-means clustering, symbolisation, or dimensionality reduction. At the end of the data processing it is easily possible to find relations between events in the data streams with the help of token tables and to enable symbolic search for repetitive patterns.

Key words: Data mining, time series, classification, clustering, sax, symbolic queries, lexical analysis, k-means.

Kurzfassung

Data-Mining ist ein Prozess der Verwendung von Methoden, um eine große Menge von Daten, die auch mit Unsicherheiten behaftet sein können, so aufzubereiten, dass der Mensch leichter Informationen davon ableiten kann. Diese Arbeit gibt einen Überblick über die im Data-Mining verwendeten Algorithmen und eine kurze Einführung in die wichtigsten davon. Unter diesen Verfahren sind jene besonders wichtig, die auf Symbolisation basieren. Diese erlauben eine vorteilhafte Visualisierung für den Menschen, sowie die automatisierte Suche mit lexikalischen Abfragen, zum Beispiel zum Finden von wiederkehrenden Mustern oder Ausnahmesituationen. Besonders die Methode der Symbolic Aggregate Approximation erlaubt eine effiziente Reduktion der Dimensionalität und Indexierung mit Hilfe von positiv semidefiniten Distanzmaßen. Nach der einführenden Übersicht wird die Anwendung auf reale Daten gezeigt, die an einer Maschine aufgenommen wurden. Zwanzig Sensoren lieferten Daten über ein Jahr lang, wobei ungefähr eine Milliarde Messwerte anfielen. Anhand von zwei Sensoren werden Pre-processing, k-means Clustering, Symbolisation und Dimensionality Reduction erklärt und angewendet. Als Ergebnis kann man Beziehungen zwischen den Datensätzen leicht finden durch Token-Tabellen und Muster erkennen durch symbolische Suche.

Schlagworte: Data-Mining, Zeitreihen, Klassifikation, Clustering, Sax, Symbolic Query, Lexikalische Analyse, k-means.

Eidesstattliche Erklärung

Hiermit versichere ich, die vorliegende Arbeit selbstsändig und unter ausschließlicher Verwendung der angegebenen Literatur und Hilfsmittel erstellt zu haben.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Leoben, 2. December. 2014 Xu Tian

Acknowledgment

- First and foremost, I thank my supervisor, O.Univ.-Prof. Dipl.-Ing. Dr.techn. Paul O'Leary who gives me the chance to complete my thesis in automation institute. I am very grateful to him for his help throughout the time when I am doing my thesis and programming.
- Herewith I also like to express to my co-supervisor, Prof. Dipl.-Ing. Dr.mont. Gerhard Rath, who has been giving me attentive advice and support from selecting the project to completing the work.
- I shall extend my thanks to Mrs. Hirtenlehner Petra and Mr. Gerold Probst for their kindness and help.
- In my daily work I have been blessed with a friendly and cheerful group of colleagues.
- I also thank my auntie Shi Xiao Fang and my parents Xu Jie Min and Shi Xiao Yuan for their financial and spiritual support during my stay in Austria.
- Last but not least, I want to thank all my friends in Leoben, for their encouragement, support and help.

Contents

Chapter 1

Introduction to Data Mining

We are living in a fast changing world, where information is exploded. To get the right information can help a doctor to diagnose his patients correctly, help an investor to get a large gain from his investment, help a machine manufacturer to produce a better functional machine, and help a shopping mall owner to stock the right commodities to satisfy his customers [13]. Concluding one can say, who gets the right information, gets the chance. To meet this demand, data mining is naturally developed.

The amount of data in data mining is tremendous and potentially infinite and often generated by real-time surveillance systems, communication networks, on-line transactions in the financial market or retail industry, electric power grids, industry production processes, scientific and engineering experiments, remote sensors, and other dynamic environments. They are temporally ordered and fast changing. It is impossible to store the entire data that flow continuously and scan it multiple times due to its huge volume. Even when a complete set of data is collected and stored in a mass storage device, it is also not economical to scan it multiple times. In addition, the raw data is usually of low level of abstraction, but most of the users are interested only in high level dynamic changes, such as trends and deviations. By finding the knowledge or patterns of the data can discover the information hidden behind the conventional events. Therefore, data mining is more and more popular nowadays.

1.1 Definitions of Technical Terms in Data Mining

Data mining : an activity that extracts some new non-trivial information contained in large data bases, data warehouses and other information bases, for the purpose of finding hidden patterns, unexpected trends or other subtle relationships in the data, and then transforms it into valid, novel, potentially useful, and understandable information in our real world. In a simple word, to find the correct and useful information in the given data set.
The raw data before mining is usually massive, uncompleted, with noise, and vague, therefore, in order to find the potential useful information which is hidden in the data and unknown in advance, a combination of techniques including machine learning, statistics and the knowledge about the data bases, and etc, is used. Various methods have been developed for data mining.
Not all the information discovery tasks are regarded as data mining [12]. For example, finding out an individual record in a data base management system, or looking for a specific web page by using a search engine of the internet, these tasks are belonged to information retrieval. Although these tasks are also important, may be related to complicated algorithms or data structures, these tasks are done by relying mainly on traditional computer science and technology, making use of the obvious features of the data to build the index structures, so that the retrieval information effectively. However, data mining technology can enhance the capacity of information retrieval systems.

1.2 Tasks of Data Mining

Prediction Task: Forecasting the specific value of a certain property (target) based on the value of other properties, such as regression, classification, anomaly detection.

Description Task: Find the potential contact modes in summarized data, such as correlation analysis, evolutionary analysis, cluster analysis, sequential pattern mining.

1. Data mining can do the following six different things (analytical method):

   - Classification
   - Estimation
   - Prediction
   - Affinity grouping or association rules
   - Clustering
   - Description and Visualization

2. Classification of data mining
   Above mentioned six data mining analysis methods can be divided into two categories: direct data mining; indirect data mining [25].
   Direct data mining: The ain is to build a model in the available dataand then use the model to describe the remaining data as a particular variable.
   Indirect data mining: Instead of selecting in targets a specific variable to be described by the model, it builds a certain relationship in all the variables.
   Classification, Estimation, Prediction are in the direct data mining catagory; Affinity grouping or association rules, Clustering, Description and Visualization are in the indirect data mining.

3. Introduction of various analytical method:

   - Classification: First, select a classified training set in the data, then using data mining technology to build on this training set

a classifying model, which can be used to classify the rest of the data.

Examples:
a. Credit card applicants who are classified as low, medium and high risk;
b.Troubleshooting:
Using data mining technology for the whole process of production to do the quality monitoring and analysis, by the means of building fault maps, real-time analysis of the causes of defects in products, that can greatly improve the rate of the products.

Note : In Classification the number of the classes is definited and predefined

- Estimation: Estimation and classification are similar, the difference between them is that classification describes the output of discrete variables, while estimation handles the output of the continuous. The number of classes of classification is a predetermined, while the amount of the estimation is uncertain.

  Example:
  a. According to its buying patterns, estimate a family's income;

  Generally speaking, the estimation is a pre-step work of classification. Given some input data, via estimation to get the value of the unknown continuous variables, and then, use a preset threshold to do the work of classification.

- Prediction: Generally speaking, prediction is connected to classification or estimation, that is, classification or estimation issues a model, which is used to predict the unknown variable.

From this sense, prediction has actually no need to be divided into a separate class. Its purpose is to predict unknown variables in future, this prediction takes time to be verified and it must be of a know accuracy after a certain time.

- Affinity grouping or association rules: Predict things that probably happen together;

  Examples:
  a. Supermarket customers while buying A, often buying B, ie. A $\Rightarrow$ B (association rules)
  b. Customers after buying A, some time later, will buy B (Sequence analysis)

- Clustering: Clustering is to divide the records into groups, by putting similar records into a cluster. The difference between classification and clustering is that clustering does not rely on the predefined class and does not need a training set.

  Examples:
  a. Some cluster specific symptoms may indicate a specific disease;

  Note : Clustering is an important step of data mining, For example, "what kind of promotions is the best welcomed by the customers?" To answer the question, the first step is to cluster the whole customers into different clusters in accordance with the customers' features, and then let the customers in different clusters answer the question, thus to get better result.

- Description and visualization: It is a task of the representation of the result of data mining;

1.3 Several Popular Data Mining Techniques

1. Decision tree:
   A decision tree is a prediction model: it represents a reflection relation between the object attribute and the object value. Each node on the tree represents a certain object, and each twig path represents a certain possible attribute value, while every leave node corresponds to the object values that the objects gets on the path from the root node to the leave node.
   A decision tree has only single output, if there is a plural output, another independent decision tree can be built to deal with the different output. Decision tree is a technology very commonly used in data mining because its high visualization, it can be used to analyse data and to make the prediction as well.
   In decision analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

2. Clustering:
   According to Vladimir Estivill-Castro, the notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms [1]. Clustering techniques are used to divide the data into groups on the basis of data similarities. It provides a mechanism which can automatically find some structures in large data set which otherwise is difficult to summarise or visualise. Different researchers employ different cluster models, and different cluster models use different algorithms. The notion of a cluster, as found by different algorithms, varies significantly in its properties. Understanding these cluster models is the key to understand the differences between the various algorithms.

3. Bayesian classification:
   Bayesian classification is the general term for a class of classification algorithms that are all based on Bayes' theorem, it is referred to as a Bayesian classifier [15].
   Bayesian classification is based on probabilistic reasoning, that is, in the situation when the various conditions are uncertain, only the probability of an occurrence is assumed, by using the mathematical probabilistic principle to complete the reasoning and decision-making tasks. Probabilistic reasoning is opposite to ascertain reasoning. The naive Bayes classifier is based on Independence assumption, that is, the features of the assumed sample is not related to that of other samples.

4. Linear regression:
   In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.
   Linear regression is a regression analysis in statistics. It builds a model of the relationship between the least square function in so called linear regression equation and its one or more independent variables or dependent variables, then based on the model to do the analysis [7].

5. Association rules (AR):
   Association rule learning is a popular and well researched method which is used to discover interesting relations between variables in large databases [29].
   It is used to identify strong rules discovered in databases using different measures of interestingness.

6. Logistic regression:

In statistics, logistic regression, or logit regression, or logit model [7], is a type of probabilistic statistical classification model [3], one of the discrete choice models. It belongs to the category of the multivariate analysis, it is a common method used in sociology, biostatistics, clinical, quantity psychology, econometrics, marketing and other statistical empirical analysis.

7. Text mining:
   Text mining is sometimes also called text exploration or text data mining, is roughly equivalent to text analysis, through which high-quality information in text is discovered.

Chapter 2

Understanding Data

Before mining the data, we must have the data prepared. To fulfill this task, we must study the attributes and the values of the data carefully. In the real world, data is usually mixed up in noise, of massive amount, or coming from heterogeneous data sources, etc. Therefore, the knowledge about the data is very useful for data mining. For example, of what type of attributes or syllables consist the data? What type of data value each attribute has? Which attributes belong to discrete, which to continuous, how the data look like? How distribute the data? Which method can be used to observe the data in a visual way? Is it possible to measure the similarities between certain objects and other objects?

2.1 Data Object and Attribute Type

2.1.1 Data Objects

To get known to data, first we must know what the data objects is. So what is the data object?
Data object is the abstract of complex information, which the software must understand. Event with only one value (for example, width) is not a data object.
Data object is a set of data elements of same properties, also regarded as a subset, is a concept of heterogeneous operation. It can be an external entity

(for example, any event that generate or use the information), general affairs (for example, report or statement), act (for example, making phone call), occurrence (for example, alarming), roles (for example, teacher or student), unit (for example, accountant division), location (for example, warehouse), or structure (for example, document) etc. In all, any entity which can be defined by a set of attributes can be regarded as a data object [12].
A data set is formed by data objects, and a data object represents an entity, usually, a data object is represented by attributes. When data objects are stored in data base, they are called data tuples, that is, the lines in the data base are corresponding to data objects, while the columns in the data base are corresponding to attributes.

2.1.2 Relation of Data Objects

The way that the data objects are connected to each other is called contacts, also called relationship. Contacts can be divided into three types:

- One to one (1:1): For example, a division has a manager, while each manager works only in one division, so that the relation of the manager and the division is one to one.
- One to more (1:N): For example, the relation between a teacher in a certain school and the courses he teaches is one to more, that is, each teacher teaches more than one courses, but each course is taught by only one teacher.
- More to more (M:N): For example, a student can learn more than one courses, and one course can be learnt by more than one students.

2.1.3 Attribute and the Types of Attribute

Attribute is the abstract description of an object. A concrete event has usually many properties and relations, and we call all the properties and relations the attributes of the event.

The event cannot be separated from attributes, because every event has its attributes and every attribute belongs to the event. A certain event is similar or different from another event, that means the attributes of the event are similar or different from that of the other.
Since the attributes of events are similar or different, there are many different event classes in the objective world. The events of same attributes form a class, the other events with different attributes form different classes respectively.
Attribute is a data syllable, represents a feature of the data object [21].
For example, apple is an event class which is composed by many individual events with same attributes. Pear is also an event class, also composed by many individual events with same attributes. Apple and pear are two different classes, because the common attributes of apple class are different from the common attributes of pear class.
The type of an attribute is decided by the accumulated values that the attribute may have. Attribute can be nominal, binary, ordinal, or numerical [9].

- Nominal attribute: The values of nominal attributes are only different names, that is, a nominal value provides only enough information to distinguish objects. For example, post code, id no. of employers, the color of eye ball, gender, etc.

- Binary attribute: Binary attribute has only two states: 0 or 1, where 0 represents that the attribute is not existed, while 1 represents that the attribute is at present. Binary attribute is also called Boolean attribute, if two status means true and false, the attribute to describe a patient can be: 1 represents he smokes, while 0 represents not.
  Binary attribute can be divided into symmetrical and asymmetrical attribute. So called symmetrical attribute means that the two states have the same value with same weight power, that is, it makes no difference weather using 0 or 1 to represent the whichever result of the two. Asymmetrical attribute means that the two states are not of same importance [35].

- Ordinal attribute: Ordinal attribute provides enough info to determine the order of the object For example, the hardness of ore (good, better, best), scores, street numbers.
- Numerical attribute: Numerical attribute can be divided into two types:
  - Interval scaling: For interval scaling, the difference between values is significant, that is, the existing measurement unit, such as, calendar date, Celsius or Fahrenheit degree.
  - Ratio scaling: As for ratio variable, the difference and ratio are all significant. For example, absolute temperature, monetary volume, counting, age, quality, length and electric flow.

Attributes have also other classified method, for example, the classifying algorithms developed in machine learning field often divide the attributes into discrete or continuous types, that means that attributes are either discrete, or continuous.

2.2 Statistics Knowledge Related to Data Mining

In order to mine the data successfully, to know and understand thoroughly the data plays a important role. The basic statistical description can be used to distinguish the properties of the data, to highlight which data values should be regarded as noise or discrete points.
Therefore, we must master some statistics knowledge. There are two fundamental statistics descriptions which can be used to describe data.

- Central tendency: Central tendency means that the data is distributed near the center or center point. Central tendency reflects that a set

of data tends to be close to a certain central value, the number of the data near the central point is big, while the number away from the central point is small. The measurements of central tendency can be mean value, medium, mode and midrange.

- Discrete distribution: Data of discrete distribution, also called scattered data, is of scattering properties and very different from each other, therefore, it is very difficult to use the mean value to represent this kind of data. In this case, some other measurement methods can be used, such as, range, inter-quartile range, variance, standard deviation, standard tolerance, and coefficient of variation, etc., of all these methods variance and standard deviation are mostly often used.

2.2.1 Measures for Central Tendency

1. Mean:
   A mean value means the central tendency measure of a set of data, which is an index reflects the central tendency of the data. The most common and efficient measurement for centering the data set is the arithmetical mean value.

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{2.1}$$

2. Weighted arithmetic mean:
   Formally, the weighted mean of a non-empty set of data $\{x_1, x_2, ...x_n\}$ with non-negative weights:

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{2.2}$$

   The weight power reflects the significance, importance and frequency of the numerical value.

3. Median:
   For sloping (asymmetrical) data, the better measure for data center is median, real numbers are ordered based on their sizes (ascending or descending):

$$Q_{\frac{1}{2}} = \begin{cases} x_{\frac{n+1}{2}}, & if\ n\ is\ odd \\ \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & if\ n\ is\ even \end{cases} \tag{2.3}$$

4. Mode:
   The mode is the value which occurs in an ensemble most frequently. Using the mode to represent a set of data is suitable to large set of data, and only when is the mode not influenced by some extreme data. The calculation of the mode is simple, a set of data can be without a mode or having several modes. In Gaussian distribution (normal distribution) the mode is the peak value.

5. Normal distribution:
   Normal distribution can also be called Gaussian distribution. If a random variable X obeys a Gaussian distribution, whose mathematical expectation is μ and square deviation is σ^2, it is noted as $N(\mu\ \sigma^2)$ The probability density function is its normal distributed expectation function μ, which decides its location, its standard difference decides the distribution amplitude. The often called standard normal distribution is the normal distribution with $\mu = 0$,$\sigma = 1$ [2].
   The probability density function curve of normal distribution is bell-shaped, so that it is also called bell curve. The normal curve is like a bell, its sides are low, its middle is high, both sides are symmetrical, the total area between the curve and its transverse axis is equal to 1.
   The normal distribution is a probability distribution, which has the continuous random variables distribution of two parameters μ and σ^2, where the first parameter μ obeys the mean value of the normal distri-

bution random variable, and σ^2 is the square deviation of this random variable. The probability rule which obeys normal distribution random variable is that the closer the values to μ are, the higher the probabilities, the farther the values are away from μ, the lower the probabilities.

2.2.2 The Absolute Index of Discrete Extent

1. Range (R): Range, also called full pitch, is the deviation between the maximum value and minimum value in a set of data.

$$R = max(X_i) - min(X_i) \tag{2.4}$$

2. Quartile deviation:
   Quartile deviation is the deviation between the third quartile and first quartile, also called internal spur or quartile range.
   Calculation formula:

$$Q_r = Q_3 - Q_1 \tag{2.5}$$

   What is quartile? We can understand in this way:
   Quartile is the points dotted at a certain distance from each other on the data distribution, which is used to divide the data into similar sized data sequences.
   2-quartile is a data point, which divides the data distribution into high and low two parts, 2-quartile is corresponding to midrange.
   4-quartile is 3 data points, which divide the data into 4 equally sized parts, making each part represent a quarter of the data distribution.
   100-quartile is called percentile, which divides the data distribution into 100 equally sized sequences.
   Quartile deviation reflects the discrete extent of the 50% data in the middle of the data distribution. The smaller the value of the quartile is, the more concentrate the data in the middle; the bigger the value is, the more discrete the data in the middle. Quartile deviation is not

influenced by extreme values, therefore, it remedies in certain extent the defect of the range.

3. Variance and standard deviation:
   Variance is the mean value of the square of the difference between actual value and the expected value, while standard deviation is the arithmetic square root of variance.

$$s^2 = \frac{1}{n}[(x_1 - x)^2 + (x_2 - x)^2 + ... + (x_n - x)^2] \tag{2.6}$$

   The computation of the variance and standard deviation is scalable in large databases [19].

2.3 Similarity and Dissimilarity of Data

Similarity and dissimilarity are very important concepts, used in various data mining techniques, such clustering, closest neighborhood classification, and anomaly detection, etc. In many cases, whenever the similarity or dissimilarity has been calculated out, the original data is no longer needed. For easy to handle sake, we use the term of proximity to describe the similarity or dissimilarity.

How to ascertain the proximity of data? First, the types of the measurement must be applicable to the types of the data. For many kinds of dense, continuous data, it is normally the distance measurements to be used, such as Euclidean distance method, etc. The differences between attributes values are to be used to describe the proximities between each of the continuous attributes. The distance measurement provides a good technique to organize the set of differences into the whole proximity measurement.

For sparse data, which usually includes asymmetrical attributes, the measure methods like cosine, Jaccard and General Jaccard are applicable.

In certain situations, in order to get the appropriate similarity measure, it is important to transform or normalize the data. This will be further discussed in next chapter of data preprocessing.

1. Distance measurement:
   There are several well known distance measurement methods used in data mining, include Minkowski, Manhattan and Euclidean, etc. Following is the brief introduction into these methods.

   - Manhattan distance:
     On a flat surface, the Manhattan distance between the point i of coordinate (x_1, y_1) and point j of coordinate (x_2,y_2) is:

     $$d(i,j) = |x_1 - x_2| + |y_1 - y_2| \tag{2.7}$$

     It must be noted that the principle of Manhattan distance is based on the turning degree of the coordinate system, instead of on the translation or mapping of the system on the coordinate axis.

   - Euclidean distance:
     the Euclidean distance between the point i of coordinate (x_1, y_1) and point j of coordinate (x_2,y_2) is:

     $$d(i,j) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2.8}$$

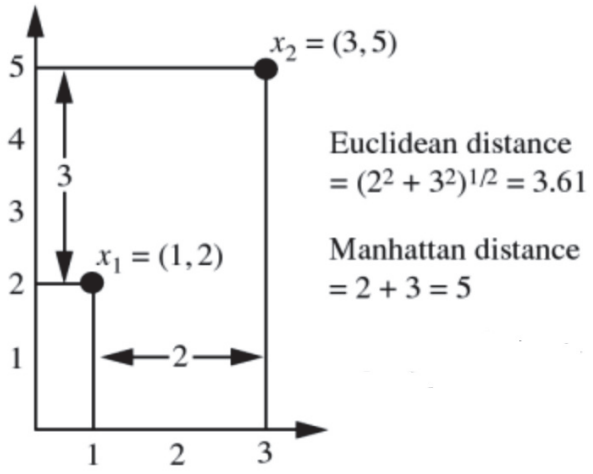

Figure 2.1: Euclidean,Manhattan distance between two object

- Minkowski distance:
  Is a generalization of the Euclidean distance and Manhattan distances.

$$d(i,j) = \sqrt[h]{|x_1 - x_2|^2 + |y_1 - y_2|^2} \tag{2.9}$$

  In which h is a real number, $h \geq 1$. When $L = 1$, it is Manhattan distance, while $L = 2$, it is Euclidean distance.

2. Cosine similarity measure:
   Usually, a document is represented by a vector, each attribute of the vector represents the occurrence frequency of a specific word(term) in the document. Of course, in real world the situation is more complicated, because the conventional words need being ignored; a same word when being dealt with by different techniques has different forms of result; different documents have different lengths and different frequencies of the word.
   Although a document has hundreds or thousands of attributes(words), the vector of it is sparse, because it has comparatively less non-zero

attribute values.
Definition of Cosine similarity:

$$cos(x, y) = \frac{x.y}{||x|||y||} \tag{2.10}$$

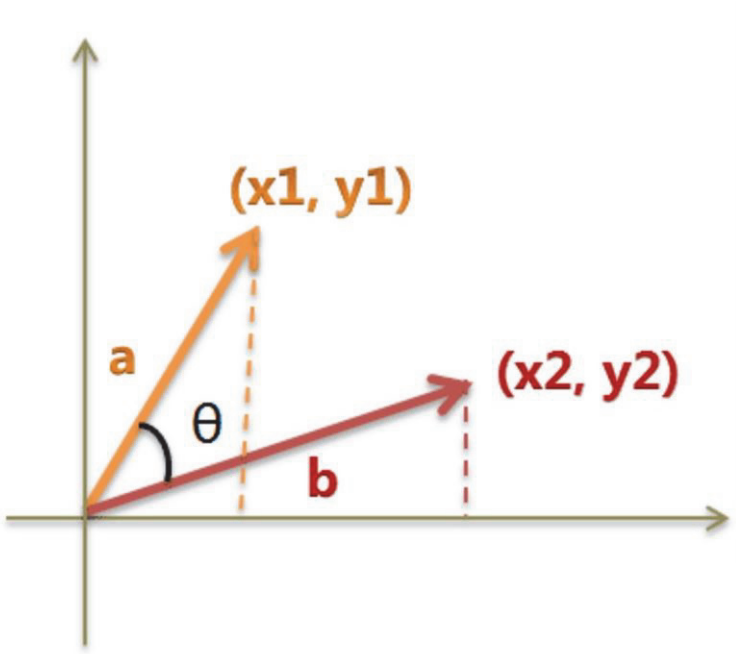

Figure 2.2: Cosine similarity

Assume x and y are two vectors of two documents, "$\cdot$" Represents the dot product of the vector, $||x||$ represents, $||x|| = \sqrt{x_1^2 + y_1^2}$. From viewpoint of concept, it is the length of the vector. Actually, the similarity of cosine is the measurement of the angular (cosine) between x and y, if the cosine similarity is 0, the angel between x and y is 90, and it means that there is no same word (term) contained in both documents. If the cosine similarity is 1, the angel between x and y is 0, that means, except for size (length), x is the same as y.

3. Jaccard coefficient:

The following is an example which helps to understand Jaccard coefficient. Assume x and y are two data objects, representing two lines of an event matrix respectively (2 events).
$X = \{1, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$,
$Y = \{0, 0, 0, 0, 0, 0, 1, 0, 0, 1\}$,
In the a.m. two lines, 1 represents that the commodity is already bought, while 0 represents that the commodity is not yet bought. The number of the commodity which are not yet bought is much bigger than the number of that already bought.
The Jaccard coefficient J is the number of matches divided by the attribute number not related to the match of f_{00}
In the above example:

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \tag{2.11}$$

in which,
the number of attributes when $f_{01} = 2$, x as 0, and y as 1
the number of attributes when $f_{10} = 1$, x as 1, and y as 0
the number of attributes when $f_{00} = 7$, x as 0, and y as 0
the number of attributes when $f_{11} = 0$, x as 1, and y as 1

Chapter 3

Data Preprocessing

What is the key in data mining? The answer is obvious: data. But not all the data is useful, because in the real world, most of the data is uneven, concept hierarchy unclear, and different quantity unit, so that it is not possible to mine it directly, or the result of the mining is meaningless. This will cause troubles in analysis of the data by leading to a false conclusion. In order to improve the result of data mining, techniques used in data preprocessing have been developed.

3.1 The Necessity of Data Preprocessing

The object in data mining is of huge amount collected from different fields or areas of the real world. In real live or production there exist many different factors which are complicated, changing, and incidentally, that cause the data we obtained might be with properties of incompleteness, with noises, in mixed and disorderly state. Hence, it is inappropriate to be mined directly. The related definitions are:

- Incompleteness:
  It means that there might be one or more values of data or its attributes missing, or some other uncertain circumstances in data.

- With noises:
  It means that there might be one or more incorrect values of data or its attributes, or exist a lot of vague information in data.

- Mixed and disorder (inconformity):
  The original data might be collected from different applied systems (say, sensor, and ect.). When these systems are not in a unified and standard management, the data collected from them will be greatly uneven. In such cases, if it is combined directly, the data will be repeated and tediously long. Therefore, the data coming from different applied systems cannot be combined directly.

Therefore, data preprocessing is a very important and very necessary step in data mining. Using preprocessing technique before data mining can greatly improve the quality of data mining, and decrease the time needed in data mining.

3.2 Main Methods for Data Preprocessing

Data preprocessing normally consists of 4 parts, data cleaning, data integration, data transformation and data reduction.

- Data cleaning:
  Data cleaning can be done by filling in the missing values, smoothing the noisy data, ascertaining or deleting the discrete points, and solving the problem of inconformity. The main purpose of it is to standardise the format, delete the deviated data, correct the faults, and cast away the repeated data. This will be further discussed in 3.3.

- Data integration:
  Data integration is the procedure of combining the data from multiple sources storing it together to build a data storehouse. This will be further discussed in 3.4.

- Data transformation:
  This is the procedure to transform the data into the data form that is appropriate to mine by the methods of smoothing processing, aggregation processing, and standardising. This will be further discussed in 3.5.

- Data reduction:
  Normally the original data is of large quantity and cost greatly in data analysis. The technique of data reduction can be used to represent the data by its greatly reduced quantity. The reduced data is much smaller but remain the approximate completeness of the original data. The result of the mining on the reduced data is the same or about the same as the result from the mining of the original data. It will be further discussed in 3.6.

It must be noted that the above mentioned methods are not independent from each other, but related to each other. For example, deleting work of tedious data is a job of data cleaning and also a job of data deduction.

3.3 Data Cleaning

Data cleaning includes 3 kinds of processing, that is, missing values processing, noises data processing and inconformity processing.

1. Missing values processing:
   For a massive database, it is quite normal to have missing values in certain dimensions or certain attributes of data which is to be analyzed. In such cases, the following methods can be used: [4]

   - Ignoring the tuple,
   - Filling in manually,
   - Using a global constant to fill in the missing value,

- Using the mean value of the attributes to fill in the missing value,
- Using the mean value of all the samples of the same type as the given tuple to fill in the missing value, or
- Using the most probable value to fill in the missing value.

Method 3-6 may causes the data incline, the data filled in may not correct. But method "Using the most probable value to fill in the missing value" is the most common method, with the help of regression, the induction based Bayesian formalism, or decision tree the estimated value can be worked out.

2. Noisy data: Just like the noises in natural world, there is also a lot of noise in data. The techniques used to get away the noises are box splitting, clustering, computer manual checking, and regression.

   - Box splitting technique:
     It is the way of dividing the data into different types, using the reasonable numerical values to replace the original data, so as to remove the noises in original data. That is to order the data first, then divide and put the data into the boxes of different depths. Then using the mean value, or median value or edge value of each box smooths the data.
   - Clustering technique:
     Clustering technique is to use the method of distinguishing the distances between the data and then dividing it into different hierarchies, so as to monitor and remove the isolated points.
   - Regression technique:
     Regression technique is to make the use of a regression model, whose predicted value is used to replace the original data. That is, a function is used to fit the data, so as to make it smooth. The methods include, linear regression, and multiple linear regression.

3. Inconformity of data:
   It is can be corrected manually by checking the documents.

3.4 Data Integration

Data integration is, as its name implies, to combine the data from multiple sources into one data store, such as data warehouse. Three points must be considered: the discovery and process of entity recognition, data tedious and the conflicts of data values.

1. Entity recognition:
   By using the way of matching the entities of real data from different sources. The real data from different sources might have different names, but might with the same attributes. During data integration, we should notice the information, such as, $stud\_Nr.$ in data A and $student\_ID$ in data B.

2. Data tediousness:
   When combining data from several databases, the tedious problem of data often occurs. The same attributes are often represented by different symbols in different database. Some tedious problems can be discovered by the relevant analysis:

$$r_{A,B} = \frac{\sum(A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B} \tag{3.1}$$

   To get away these tedious values can increase the speed and quality of the result of data mining.

3. Monitor and solve the conflicts of data values:
   In real world, the attributes of a same entity are might be different while coming from different data sources. The reasons cause this situation can be different representations of data, different measurement unit, etc.

3.5 Data Transformation

Data transformation is an inevitable result of data processing, it is mainly through the ways of smoothing, aggregation, data generalization, data stan-

dardization and attribute structuring to fulfill the job.

- Smoothing: means to remove the noises in data (specific methods include box splitting, clustering and regression).
- Standardization: means to put the data into a very small and specified space by the method of zooming the data proportionally.
- Attributes structuring: means to structure new attributes on the base of the already existed attributes, and then add them info attributes set, for the purpose of understanding the high dimensional data structure more precisely.

3.6 Data Reduction

When dealing with a huge amount of data from massive scaled databases, if mining and analyzing all the data, it will be a very huge project. This will cost a very long time and high expenses. If we can extract the main data from it, we can do the analyzing work much quicker. Such extraction techniques include generally data aggregation, dimensionally reduction (by monitoring and deleting unrelated, weakly correlated or tedious attributes or dimensions), data compression (wavelet or Fourier transform or main contents analysis), numerical value reduction (using substitute, smaller data to replace or estimate data). Main methods are regression, histogram, clustering, sampling and concept hierarchical dividing, etc. [20]
For small or medium sized data sets, the general preprocessing procedure is enough. But for a really huge sized data set, before using the techniques of data mining, an intermediate and extra step of data reduction is recommended.

Chapter 4

Time Series Data Mining (TSDM)

One kind of data sets is called time series, between whose data a relation of time exists. When time series is being mined, the relation of time in the data set must be considered. Koogh [6] thinks that time series are everywhere, for example, image data, text data, photo data, handwriting data, and brain scanning data etc. all can be regarded as time series. It is of important theoretical value and practical significance to study the methods of time series mining, with which one can get useful information from huge and complicated time series data efficiently.

4.1 Time Series and Its Application

Time series data mining is to study the time features of the information, to get know detailed evolutionary mechanism of the things. It is a efficient way to get useful knowledge of the things.

- In a statistical sense, time series means that the values issued by a certain index at different times are ordered in accordance with their issuing times chronologically.
  Time series mining includes analysis of the objectively recorded past

behaviors of the things to find the inherent regulations of the things, so as to predict the future behaviors of the things, and to fulfill other decisive work.
In simple words, time series data mining is the work to extract from huge amount of time series data the information and knowledge which is related with its attribute of time, unknown in advance, but potentially useful and use it for short, medium and long term prediction of the behaviors of the society, economy, military and life etc.

- In mathematical sense, if we observe or measure a certain variance of a certain procedure, in a series of time $t_1, t_2 ... t_n$ to get the discrete ordinal set $X_{t1}, X_{t2}, ... X_{tn}$, it is called a discrete numerical time series. Assuming $X(t)$ is a random procedure, $X_{ti} (i = 1, 2, ... n)$ is called a sample realization, that is, a time series.

The study of time series must be done in accordance of appropriate theory and techniques, the various kinds of time series indicate that the study of time series must be associated with the features of different kinds of time series, so as to find the appropriate methods to build the model.
Time series can be divided as follows:

- Univariate time series:
  For example: the number of the commodity for sale. The regulated information of such time series can be obtained from its variable process of a single variable.

- Multivariate time series:
  For example: weather data such as, temperatures, pressures, rainfall, etc. Data mining of this kind of time series need to discover the relationship between each variable.

- Discrete time series:
  If in a series the time parameters correspondent to every sequential value are discontinuous points, such time series is called discrete time series.

- Continuous time series:
  If in a series the time parameters corresponding to every sequential value are continuous functions, such time series is called continuous time series.

The distribution regulation of series: The statistical features of series can be smooth or fluctuate regularly, and this is the basis of analysis. Besides, if series is distributed in accordance with a certain regulation (say, Gaussian distribution), it is the theoretical basis of the analysis of time series.

4.2 The Main Research Contents of Time Series Data Mining

4.2.1 Time Series Data Transformation

Time series is normally very massive, mixed with noise data, and with missing values. It is not economic or not possible to mine it directly. Therefore, before mining the time series data, it is necessary to transform the original data to enable mining.

Transformation of time series data means to map the original time series into a certain feature space, then use its mapping in this feature space to describe the original time series. In this way, the data is compressed and the costs occurred in calculation can be reduced.

The already existed representation methods are mainly discrete Fourier transform, DFT [26], singular value decomposition, SVD [6] , discrete wavelet transform, DWT [33], dynamic time warping, DTW [5], piecewise aggregate approximation, PAA [6], piecewise linear representation, PLR, and piecewise polynomial representation, PPR, etc.

4.2.2 Time Series Prediction

An important application of time series analysis is prediction, i.e., on the basis of the past variation characteristics and trend of the data to predict

the future value of its attributes. The main prediction methods are divided into three kinds: deterministic time series prediction, random time series prediction, and other time series prediction.

- Deterministic time series prediction:
  For time series of smooth variation characteristics, it is feasible to assume that its future behaviors is related to its present behaviors, and to use the present values of its attributes to predict the future values of its attributes.
  Definition:

  - Long term trend: the probabilities that a value will be steadily increasing, reducing or remaining with the changes of time.
  - Seasoning change: periodical change patterns in a certain period of time.
  - Random changes: changes caused by uncontrollable accidental factors.

  Assuming T_r is long term trend, S_t is seasoning change, R_t is random changes, and Y_t is the observing records of the observing object. The commonly deterministic time series prediction models are:

  - Adding model: $Y_t = T_t + S_t + R_t$
  - Multiple model: $Y_t = T_t \cdot S_t \cdot R_t$
  - Mixed model: $Y_t = T_t \cdot S_t + R_t$ or $Y_t = S_t + T_t \cdot R_t$

- Random time series prediction methods:
  This kind of prediction methods, especially the chaotic time series prediction [14] and the time series based on neural network prediction [34], are mainly of using an embedded space or using a neural network.

- Other time series prediction techniques:
  Other techniques like sliding window two times auto regression model, time series prediction based cloud model, etc.

4.2.3 Similarities Searching in Time Series Database

The searching of similarities of time series is the searching of the contents. Since time series is continuous and fluctuate, some special problems occurred in the searching of its similarities.
First, how to definite the similarity? For easier sake, we give several symbols to represent the time series and its similarities.

- $X = \{x_t | t = 1, 2, , ...n\}$ represents a series
- $Len(X)$ represents the length of the series
- $First(X)$ represents the first element of series X
- $Last(X)$ represents the last element of series X
- $X(i)$ represents the value of X at the time of i
- the "$<$" relation between elements in series, in series X, if $i < j$, then $X(i) < X(j)$
- subsequence represented by X_s, if series X has K subsequences, these subsequences are represented by $X_{s1}, X_{s2}...X_{sk}$
- the relation of "$<$" between subsequences, X_{si} and X_{sj} are subsequences of X, if $First(X_{si}) < First(X_{sj})$, then $X_{si} < X_{sj}$
- overlap of subsequences, assuming X_{s1} and X_{s2} are two subsequences of X, if $First(X_{s1}) \leq First(X_{s2}) \leq Last(X_{s1})$ or $First(X_{s2}) \leq First(X_{s1}) \leq Last(X_{s2})$ establishes, then X_{s1} and X_{s2} is overlapped.

Usually, similarity matching can be divided into two kinds:

- Whole matching:
  Given N number of series $Y_1, Y_2, ..., Y_n$ and an inquiry series X, all of them having the same length, if $D(X, Y_i) < \varepsilon$, we say that X and Y_i are whole matching. In which D is one kind of distance measurement. ε is a predetermined value.

- Subsequence matching:
  Given N number of series of different length $Y_1, Y_2, ..., Y_n$, an inquiry series X and a parameter ε, the subsequence matching is a job to find a certain subsequence in $Y_i (1 \leq i \leq N)$, to make the distance between this subsequence and $X \leq \varepsilon$

At present, the methods for measuring the similarities of time series are mainly based on distance measurement, including Euclidean distance [13], DTW [5] etc. The speed of similarity searching by measurement based on Euclidean distance is faster than that based on DTW distance, while the result from DTW searching is better than that from Euclidean method. The lower bounding function can be associated with DTW method to accelerate the searching speed.

4.2.4 Visualization of Time Series

Visualization mining of time series is a comparatively new area of data mining research, which is also a very prospect research area of potential applications. The visualization mining of time series is to make the use of image techniques, virtual reality technology and data mining techniques to display the complicated time series in the way of understandable, visual graphics. The visualization of time series is a research direction which has a prospect of application [16]. The related methods have been developed include time series spirals, time searcher, vizTree and time series bitmaps, etc.

SAX Representation

Symbolic representation of time series has attracted much attention recently, because by using this method we can not only reduce the dimensionality of time series, but also benefit from the numerous algorithms used in bioinformatics and text datamining.

The symbolic aggregate approximation method (SAX) is the most powerful one in time series data mining.

SAX transforms a numerical time series into a sequence of symbols, by representing the values in the time series by a symbol of a finite alphabet. This method is very simple and does not require any prior information about the

time series flowing in computer system (except that the distribution must be Gaussian with zero mean and unit variance). SAX representation is basically consisted of the following steps:

- Raw data pre-processing,
- Divide a time series into segments of length L (clustering),
- Choosing the breakpoints,
- Symbolization of time series.

In chapter 6 we will describe in detail the process of using sax to deal with two exemplary time series of tonnage and slew collected from a real production system.

4.2.5 Segmentation and Model Discovery of Time Series

Model discovery is one of the important research contents of time series data mining, in which a large number of achievements are received. For different application purposes, the models are to be discovered in time series database also different, say: specific model, frequent model, periodical model, interested model, surprise model, anomaly model and exceptional model, etc. In order to get a model from time series, a certain algorithm is needed to segment a long time series into a certain number of comparatively shorter subsequences, so as to analyze these subsequences by classifying or clustering way, monitor the variable points in time series, build a dynamic model for the segmented time series [8].
There are two main applications of time series segmentation, that is: 1. to monitor system model changes, i.e., when the model or parameter of the system, which generate the time series, changes, the segmentation algorithm can monitor the time at which the changes happen. 2. use segmentation algorithm to build the advanced data representation of time series, so as to

index, cluster and classify the time series [6]. Therefore, time series segmentation research has a important theoretical value and realistic significance, which has become a main research content of time series data mining.

Chapter 5

Time Series Data Clustering

Clustering is one of the main tasks in data mining, in which data objects are divided into several subsequences which are also called clusters. The dividing work is done on the basis of the similarity of the data objects, that means, objects which are similar to each other are being put in the same cluster.
In cluster analysis, most similar data objects are discovered on the basis of some criteria for comparisons. Clustering aims to increase the efficiency of similarity among members in a cluster [17]. In the clustering domain,Han et al. [11] propose clustering method categorizations to arrange various static data.
When the features of data don't change with time, or the changes can be ignored, the data are regarded as static. The clustering methods are divided into five categories, with the names of partitioning, hierarchical, density-based, grid-based, and model-based.
The analysis of huge amount of the time series data demands the technique of pattern discovery, which is also called time series clustering. Time series clustering can be applied to many different fields, such as, e-commerce, outlier detection, speech recognition, biological system, DNA optimization and text mining.

5.1 Hierarchical Methods

One of the general clustering algorithm is hierarchical clustering, which has a powerful visualization compared with other clustering approaches [18]. Hierarchical method is to divide the given data collection into different hierarchies. It has two types, agglomerative and divisive.

- Agglomerative:
  Agglomerative is a bottom-up structure, in which each object begins as an independent group, then it merges itself with the closest object or group, the same work continues until all the groups have merged together into one group(the highest hierarchy), or until an ending condition is satisfied. Most of hierarchical clustering methods are of agglomerative structure, the difference is only in the definition of the distances.

- Divisive:
  Divisive is a up down structure, in which all the objects are being put into one cluster in the beginning, then in each step of iteration, the cluster is divided into smaller clusters, the same work continues until every object stays in one cluster, or until an ending condition is satisfied. For example, a wanted number of clusters is fulfilled, or the threshold value of distance between clusters is realized.

The basic steps of hierarchical clustering are as follows:

1. The distance between all objects are calculated and stored in a distance matrix.

2. Search in the matrix for two most similar clusters/objects.

3. Two similar clusters are merged into one cluster which has at least 2 objects.

4. The distance between the new cluster and all the other clusters are calculated and stored in the matrix

5. Step 2 is repeated until all the objects in one cluster or until an ending condition is satisfied. The advantage of hierarchical clustering is the powerful visualization, but because of its quadratic calculations complexity, this method is not applicable to medium or large sized data set. By using together with other clustering method, hierarchical clustering can be more applicable.

5.2 Partitioning Clustering

Given a data base which contain data objects n, and a wanted number k of clusters which are going to be generated, an algorithm for class partition divides the data objects into k parts ($k \leq n$), in which each part is a cluster. Normally a criteria for partitioning is taken (called similarity function), such as, distance, which is used to ascertain the objects in a cluster is similar or dissimilar. Clustering partition needs to meet 2 conditions:

- in each cluster there must be at least one object.
- each object must belong to only one cluster. It should be noted, in some fuzzy partitioning methods, the 2.$^{\text{nd}}$ condition can be loosed.

The partitioning method is to establish an essential partition first, then using the technique of iterative relocation try to improve the partition by moving the objects in different partitioning parts. The principle of a good partitioning is that, the distance between objects in a same part is the smaller the better, and the distance between objects in different parts is the bigger the better.
In order to find the best partitioning method, it is quite often that all the possible partitioning methods are required to be listed. But in actual, only two heuristic methods are mostly to be put into applications, that is:

1. k-means:
   In this method every cluster is represented by the mean value of the cluster.

2. k-medoids:
   In this method, every cluster is represented by the object which is closest to the center of the cluster.

This heuristic clustering methods are very suitable for small and medium sized data base, to be used to discover the spherical shaped clusters. For large scaled data set with complex shapes, the partitioning methods are required to be further developed.
Example of k-means algorithm:

- Input: the number of clusters k, and the dataset of objects k
- Output: clusters k
- Method:
  1. selecting by random any objects k as the center of the essential cluster.
  2. repeat
     - based on its distance to the center, give every object its closest cluster;
     - recalculating the mean value of each cluster;
     - until no changes will occur;

5.3 Density-Based Clustering, Grid-Based Methods and Model-Based Methods

5.3.1 Density-Based Clustering

Most of the cluster partitioning methods are based on the distance between objects. Such methods can find only spherical clusters, having difficulties to find other shaped clusters. Therefore, the clustering methods based on

density have been developed to meet the demand. The main idea of density based clustering is as long as the density of the objects or data in neighborhood has reached the threshold value, the clustering work continues, that is, every data in the given class must at least obtain a certain given points in a given area. Such methods can be used to filter noises data, to discover clusters of any shapes.
DBSCAN [17] is a typical density-based clustering method. It uses a threshold value of the density to control the growing of the clusters. While OPTICS [17] is an another density-based clustering method, which calculates a cluster sequence, then the sequence do the clustering analysis automatically and interactively.

5.3.2 Grid-Based Methods

By this method the object space is quantized into limited number of units, forming a grid structure. All the clustering work is done within this grid structure(quantized space).
The main advantage of this method is its fast speed. The time needed in calculations is independent from the number of the data objects, but related to the number of the units on each dimension in the quantized space. A typical grid-based method is STING, while CLIQUE and WaveCluster [17] are regarded as both grid-based and density-based methods.

5.3.3 Model-Based Methods

In this method, it is assumed that every cluster has a model. It puts the data into different modeled clusters by looking for the data which is the best matching of the given model. A model-based algorithm can locate the clusters by building a density function which reflects the data points space distribution. It can automatically decide the number of the clusters on the basis of standard statistics number, consider the noise data and isolated data, so as to produce a very healthy clustering method.

5.4 Summary

1. Partitioning methods
    - To discover spherical clusters which exclusive to each other;
    - Distance based;
    - Mean value or center point are used to represent the center of the cluster;
    - valid for small and medium sized database;
2. Hierarchical methods
    - Other techniques can be integrated;
    - The errors in combining or in dividing cannot be revised;
3. Density-Based Clustering
    - To discover clusters with any kind of shape;
    - Can be used to filter out noise data;
4. Grid-based methods
    - Using a multi-resolution grid-based data structure;
    - Fast processing (independent from the number of data objects but depending on the size of the net work);

Chapter 6

Application Example

This chapter will explain in detail the process through which useful and interesting information is mined out from two sets of real data "tonnage" and "slew".
"Tonnage" and "slew" are two exemplary sets of data taken from a huge amount of measurements data set, which is recorded by a monitoring system of a real production process consisted of twenty sensors and lasted for more than one year.
We make the use of some algorithms of data mining, such as, preprocessing, clustering (K-mean), symbolization, dimensionality reducing and visualization and so on, and via MATLAB to realize the data mining job, so as to discover useful and interesting information, that is, by combining the useful information contented in the two sets of data we obtain the working information of the machine being monitored during the production process.

6.1 Preprocessing of Data

"Tonnage" and "slew" are two sets of time series data collected from a monitoring system of a machine for 24 hours or 1440 minutes or 86400 seconds, collecting frequency is one signal at each second. That means from the 86400 seconds of real time monitoring we get 2 sets of time series data of length

86400, called "tonnage" and "slew" respectively. Both sets of time series are of datatype <86400·1 double>.

- Definition of time series data
  Time series is a series of values or elements issued or recorded by a certain phenomenon or index at different times and ordered chronologically, written as $X = \{x_1(v_1, t_1), x2(v_2, t_2)...x_n(v_n, t_n)\}$, and element $x_i = (v_i, t_i)$ represent that the recorded value of time series is v_i at time t_i, and the recorded time t_i is strictly added as $(i < j \Rightarrow t_i < t_j)$. Normally, the sampling time intervals of time series are equal, that is $\triangle\ t = t_{i-1} - t_i$, which can also be regarded as $t_1 = 0$, $\triangle\ t = 1$, in such case, time series $X = \{x_1(v_1, t_1), x_2(v_2, t_2)...x_n(v_n, t_n)\}$can be simplified as $X = \{x_1, x_2...x_n\}$ [32].

"Tonnage" is the time series which records the moving speed of the monitored machine in real production process for a whole day. "Slew" is the time series which records the moving directions of the machine in a whole day production process. Time is a very important attribute of time series. The useful information obtained from the mining of time series is dependent on time, for example, the operation situations of the machine at a certain time point or at several time points, the abnormal operation situations at a certain or at several time points, or some operation situations last for a certain time period, etc, are totally dependent on time.
In the mean time,we found "tonnage" and "slew" are also time stamped. That means, this category of the temporal data has explicit time related information. Relationship can be quantitative i.e. we can find the exact temporal distance between data element [22].

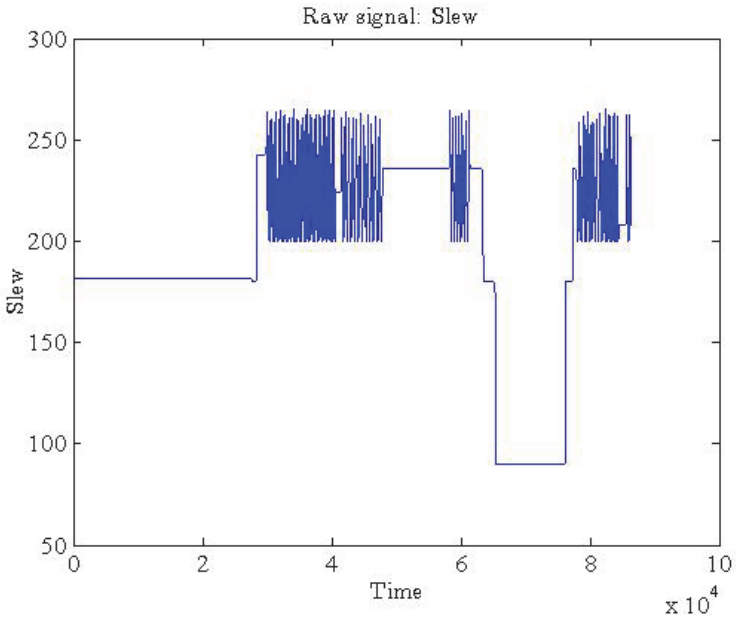

Figure 6.1: Data of "slew"

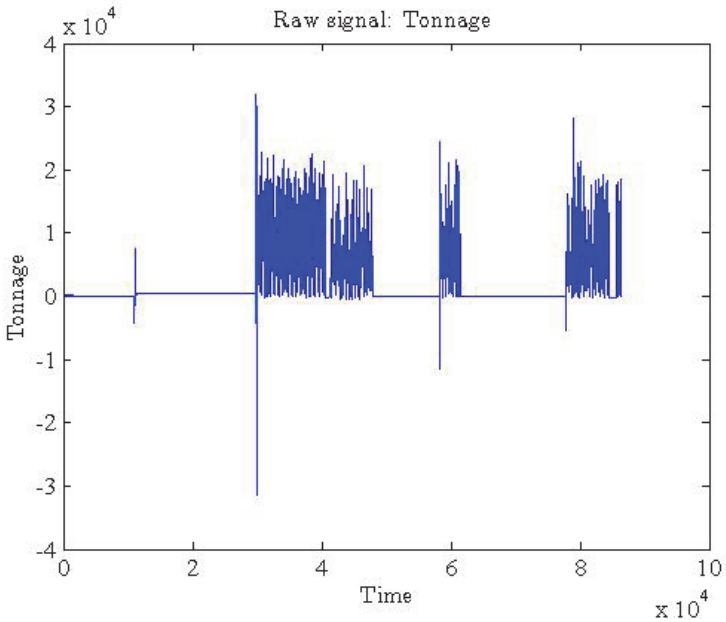

Figure 6.2: Data of "tonnage"

Fig.6.1 and Fig.6.2 show the original data "tonnage" and "slew" respectively, in which coordinate X represents time, while coordinate Y represents the true value of tonnage or slew at a certain time point.
In Matlab workspace we see that the maximum and minimum values of "tonnage" and "slew" are both "NaN", which means that there are some missing values in these two sets of data.
Hereunder we make time series data of "tonnage" as an example:
By searching we found that the values at 63 time positions from 1341 to 1403 are empty, that means there are 63 missing values in the whole time series data of "tonnage". The reasons of causing such problems are different, for example, equipment failure while collecting the data, signals from malfunctioned sensor, or mistakes made by man or machine while recording the data, etc. If we do not deal with these missing values first, there will be various abnormal detections in next steps of the mining work.
As mentioned in chapter 3, missing values are very common in real data sets, because in real production there are various complicated, changeable and accidentally factors, which cause the original real data sets incomplete.
There are 6 methods which can be used to deal with the missing values: 1.ignoring the tuple, 2.filling in by manual, 3.using a global constant to fill in the missing value, 4.using the mean value of the attributes to fill in the missing value, 5.using the mean value of all the samples of the same type as the given tuple to fill in the missing value, 6.using the most probable value to fill in the missing value. While method 6 is the most often used one up to now. It makes use of most of the already existed information to estimate the missing values.
The missing values in "tonnage" are continuous, that means, only in the time period from 1341 to 1403 loss values. Therefore, we can use a very simple method to estimate the missing values. Linear estimating is the method, that is, using the value before the missing value period and the value after the missing value period to linear estimate all the missing values in this time period.We can do this in MATLAB:
MATlAB code:

```
Inds=find(isnan(tonnage));
```

```
n=length(Inds);
a=linspace(tonnage(Inds(1)-1),tonnage(Inds(n)+1),n);
```

In which n means the length of the time period,in which values are missing, $Inds$ is the time position where value is missing, the values stored in a are to be used to fill in missing values.
Function: create a vector of n linearly spaced between and including the values before and after the time period in which values are missing, and use these values to fill in the empty points. We can see in MATLAB workspace that the data of "tonnage" after preprocessing is 86400 long, with datatype is "double", and the minimum value is -31494, while the maximum value is 32001, furthermore, there is no missing value in the time series any more.
The original time series of "slew" have also 13 missing values at time points from 1394 to 1406. We deal with them in the same way.
MATlAB code:

```
Inds=find(isnan(slew));
n=length(Inds);
a=linspace(slew(Inds(1)-1),slew(Inds(n)+1),n);
```

After preprocessing, we obtain the time series of 86400, too, with same datatype "double", the minimum value is 90, while the maximum value is 265.1, and there is no missing values either.

6.2 Clustering and Symbolization of Time Series

One of the primary tasks of data mining is clustering, whose function groups similar objects into a cluster. Clustering is the most prevalent task of statistical data analysis in various aspects. In cluster analysis, most similar data objects are discovered on the basis of some criteria for comparisons. Clustering aims to increase the efficiency of similarity among members in a cluster [17].

6.2.1 Clustering of Time Series "tonnage"

First, let us observe the distribution of time series data by histogram.

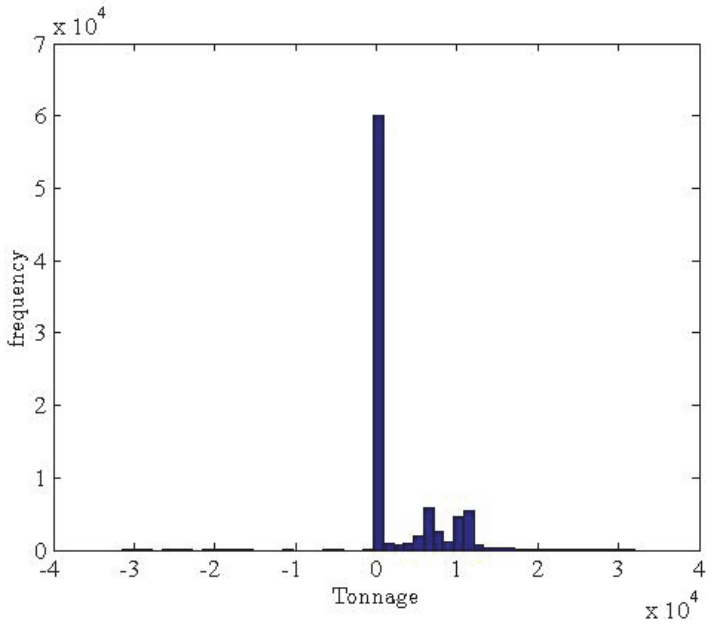

Figure 6.3: Histogram of data "tonnage"

In Fig.6.3 is to use matlab order: hist which a histogram shows the distribution of tonnage data values, to demonstrate the occurence status of each data value of tonnage at different time points .
"Tonnage" is the records of the speed of the machine, therefore, minus value is meaningless and can be regarded as abnormal values. In this case, we can easily determine the first cluster of tonnage, to cluster all the values which is less than 0 into cluster "0".
The next step is to cluster the rest data values of "tonnage". First we divide all the values from 0 to the maximum value of "tonnage" into 51 districts, in this way to observe the distribution status of all the values which are bigger than 0 in "tonnage". See Fig.6.4 .

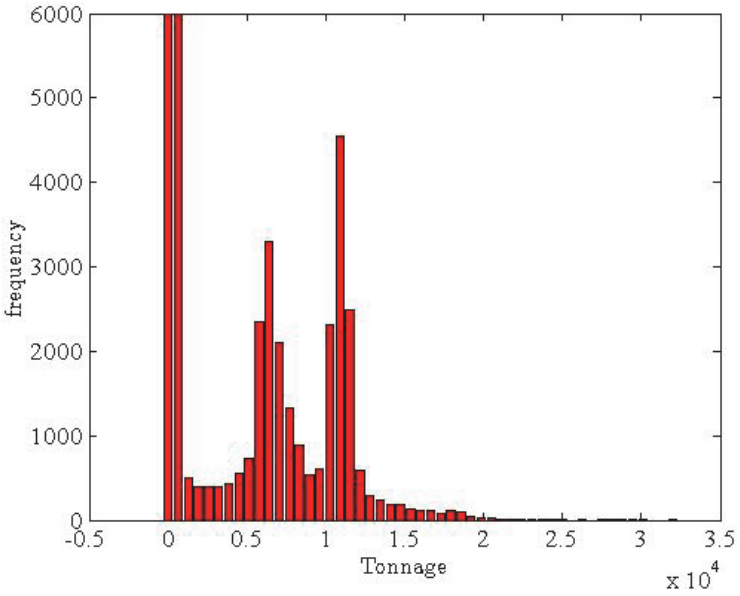

Figure 6.4: Histogram of data bigger than 0 in "tonnage"

Fig.6.4 demonstrates the distribution status of data values,which can be regarded approximately as 3 normal distributions, as shown in Fig. 6.5. Via MATLAB programming we can calculate out 2 parameters of normal distribution 1: mean value is 0, while standard deviation is 335.079. 2 parameters of normal distribution 2 are 6382.7 for mean value, 1185.5 for standard deviation. 2 parameters of normal distribution 3 are 11487 for mean value, 2008.7 for standard deviation.

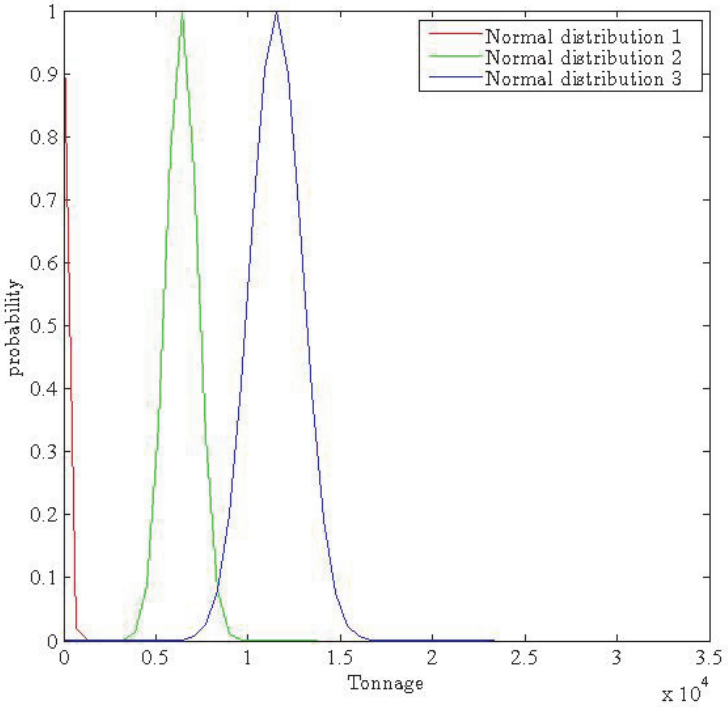

Figure 6.5: Three normal distributions for data "tonnage"

Thus, we have divided all the data which is more than 0 in "tonnage" into 3 cluster on the basis of its normal distribution status. They are cluster 1, cluster 2, and cluster 3 respectively. see Fig.6.6

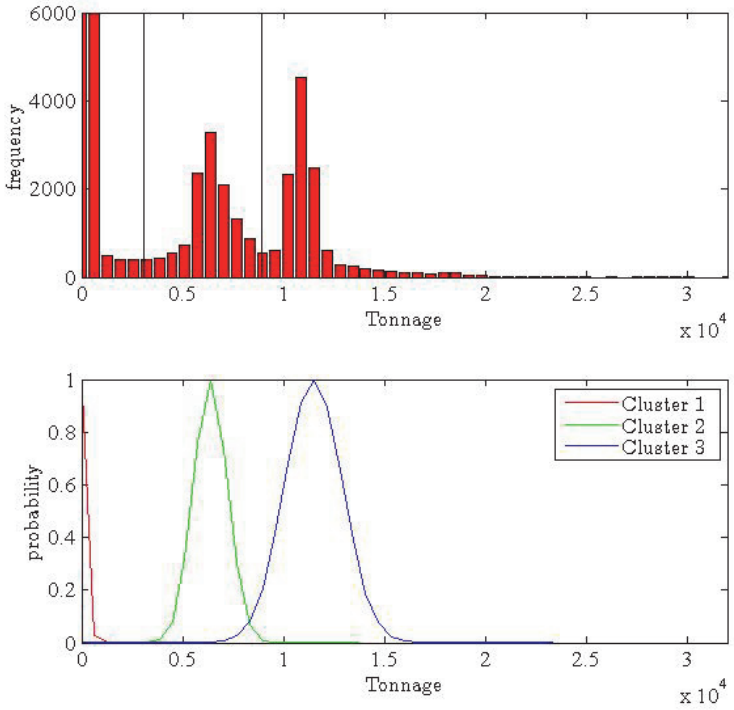

Figure 6.6: Three clusters for the data bigger than 0 in "tonnage"

By such clustering, we have obtained 4 clusters in the end. The real meanings of these clusters are that, cluster 0 contains the abnormal values of the signal,cluster 1 represents the machine in static status or the speed of the machine is 0, cluster 2 and cluster 3 mean that the machine is moving with two different speeds. In another word, through the way of clustering, we know that the machine being monitored have 4 status, that is, abnormal signal appeared, stop moving, and moving with two different speeds.

index	1	2	3	.....	86399	86400
value	75	75	70	.....	738316	8190
cluster	1	1	1	.....	2	2

Table 6.1: "tonnage"after the clustering

After the clustering work, we get a new dataset as shown in table 6.1.

The first line of this table is the index of time attribute, 2nd line is the values of tonnage, and 3rd line displays the cluster to which the value belong.

6.2.2 Clustering of Time Series "slew"

"Slew" is the records of the moving directions of the machine. Clustering of time series "slew" is different from that of tonnage. We do not use a histogram to observe the distribution status of the data of "slew", instead we want to observe the distribution status of data $\frac{d(slew)}{dt}$. So, how to get d(slew)/dt ?

First we must learn some thing about Savitzky-Golay filter, a Savitzky-Golay filter is a digital filter that can be applied to a set of digital data points for the purpose of smoothing the data, that is, to increase the signal-to-noise ratio without greatly distorting the signal.

This is achieved, in a process known as convolution, by fitting successive sub-sets of adjacent data points with a low-degree polynomial by the method of linear least squares .

When the data points are equally spaced an analytical solution to the least-squares equations can be found, in the form of a single set of "convolution coefficients" that can be applied to all data sub-sets, to give estimates of the smoothed signal, (or derivatives of the smoothed signal) at the central point of each sub-set.

The method, based on established mathematical procedures [31][10], was popularized by Abraham Savitzky and Marcel J. E. Golay who published tables of convolution coefficients for various polynomials and sub-set sizes in 1964 [28][27]. Some errors in the tables have been corrected [30]. The method has been extended for the treatment of 2- and 3-dimensional data.

In the real data processing,$d(slew)/dt = D * slew$. We can use a MATLAB function: $dopDiffLocal.m$ [23] to compute a local polynomial differential approximator. The purpose of this function is that it generates a global matrix operator which implements the computation of local differentials where the vector of $slew$ values may be irregularly spaced.

MATlAB code:

```
D=dopDifflocal(slew,Ls,noBfs);
dslew=localConv(slew ,D);
```

In this MATLAB code, D is a local polynomial differential approximator, slew is a data set having been already preprocessed, Ls is the support length used for the local differential, $noBfs$ is the number of basis functions to be used. In our program we take $Ls = 31$ and $noBfs = 3$.
Since "slew" is a time series of 86400 long, the D(local polynomial differential) having been calculated out is a matrix of 31 by 31, if we do the convolution calculation directly in MATLAB right now, failure of "out of memory" will definitely occur. In order to avoid such failure, we need another MATLAB function *localConv.m* [24] to achieve. This function *localConv.m* [24] performs local processing of a long signal using convolution. The start and end points are then corrected to ensure a complete computation over the full length of the sequence. This is related to local polynomial approximation with correct end point computation.
In which *slew* is the sequence to be processed ,D is the matrix defining the local operator.
Now let us use a histogram to observe the distribution status of data set $d(slew)/dt$ as shown in Fig.6.7.

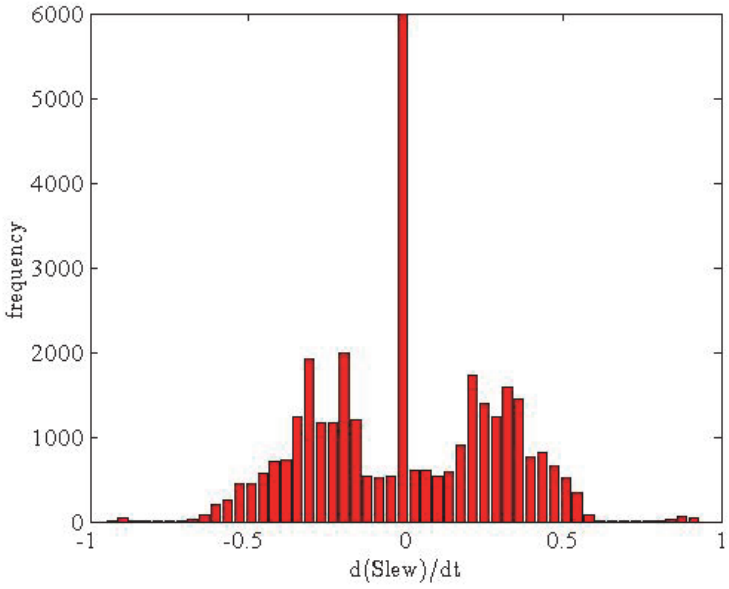

Figure 6.7: Histogram of data "$d(slew)/dt$"

Just as we have done in processing of tonnage, we can regard the distribution of data in Fig.6.7. approximately as 3 normal distributions as shown in Fig.6.8.

Via MATLAB programming we calculate out 2 parameters of normal distribution 1, that is, -0.3031 for mean value and 0.1335 for standard deviation respectively. 2 parameters of normal distribution 2 is 0 for mean value and 0.0103 for standard deviation respectively. And 2 parameters of normal distribution 3 is 0.3152 for mean value, and 0.1285 for standard deviation.

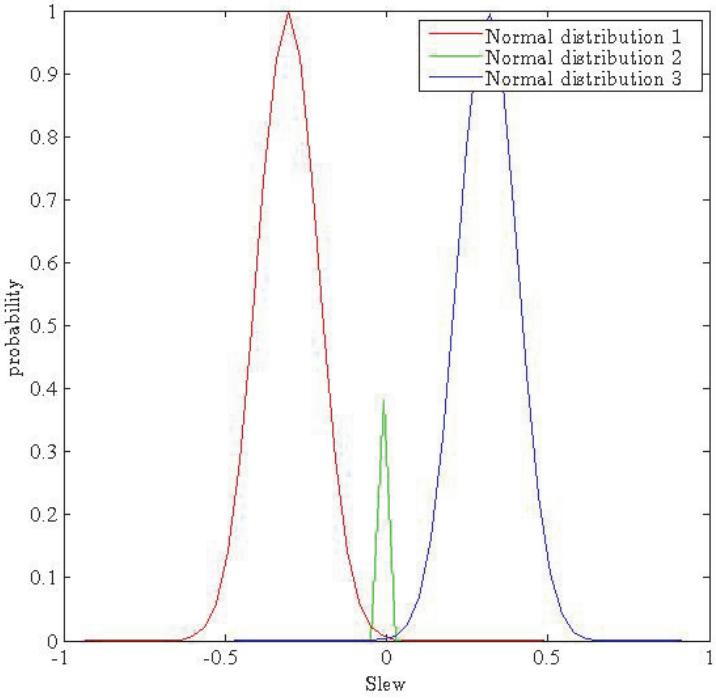

Figure 6.8: Three normal distributions for data "$d(slew)/dt$"

We divide all the data in $d(slew)/dt$ into 3 clusters on the basis of the normal distributions. They are cluster 1,cluster 2,cluster 3 respectively. See Fig.6.9.

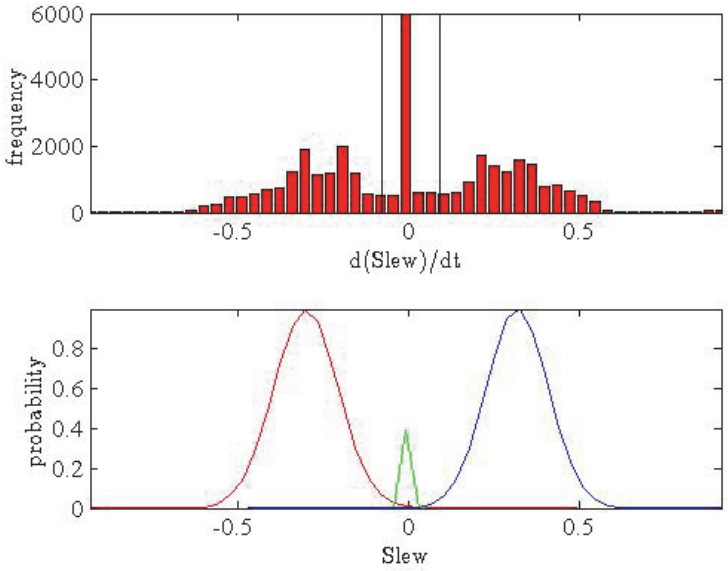

Figure 6.9: Three clusters of data "$d(slew)/dt$"

By clustering we have obtained 3 clusters, whose realistic meanings are: cluster 1 represents that the machine is moving in a direction, or can be said in left direction; cluster 2 represents the machine is in static,or means that the machine has no moving direction; cluster 3 represents that the machine is moving reversely,or can be said in right direction. In other word, through the way of clustering we have learnt that the machine is either in a state of moving to and fro or in a state of staying in static position.
When the clustering job is fulfilled, we get a new dataset as shown in table 6.2.

index	1	2	.....	86400
value	$3.1068 \cdot 10^{-15}$	$3.5527 \cdot 1010^{-15}$	.....	$7.1054 \cdot 1010^{-15}$
cluster	2	2	.....	1

Table 6.2: "$d(slew)/dt$"after the clustering

It is as the same as in table 6.1, the first line of the table is the index of time attribute, 2nd line is the values of d(slew)/dt, 3rd line displays the

cluster to which the value belongs.

6.2.3 Symbolization of Time Series Dataset

Symbolization representation of time series is a method of discretization of time series which was proposed in last decades. Because of its discretization and non real number representation features, it attracts more and more attentions in data mining field. The main idea of this method is to transform the time series represented by numerical values into symbol sequences represented by discreted symbols on the basis of a certain rule of the changes happened in the time series which is being processed.
The 3rd lines in both table 6.1 and table 6.2 are symbol sequences represented by discrete symbols, which have been transformed on the basis of the above mentioned clustering rules from the time series of "tonnage" and "slew" represented by numerical values. The tonnage dataset after the transformation is still 86400 long, but we need only 4 symbols of 0, 1, 2, 3 to represent the whole "tonnage" dataset. We can prove this point of view by the information in MATLAB workspace, where we can see the maximum value of "tonnage" after symbolic transformation is 3, and the minimum value is 0. And the same is with time series of "slew", after symbolic transformation, the length of the time series is still 86400, but we need only 3 symbols to represent the whole series data of "slew".

6.3 Dimensionality Reduction

After symbolization, the time series "tonnage" and "slew" can be represented by 4 and 3 symbols respectively, but the lengths of them are still 86400. See table 6.3 and table 6.4

index	1	2	3	.....	86399	86400
cluster	1	1	1	.....	2	2

Table 6.3: time series “tonnage” which is already symbolized

index	1	2	3	.....	86399	86400
cluster	2	2	2	.....	1	1

Table 6.4: time series “slew” which is already symbolized

The first lines of table 6.3 and 6.4 display the index of time, the 2nd lines of both tables means the cluster to which the time point belongs.
To this stage, the time series data is still not enough optimized in storage space and calculation speed for data mining, so that we need dimensionality reduction to do the further optimization.
An efficient dimensionality reduction can detect the structures and connections hidden in the original data, so as not only to eliminate the redundancy of the data, to simplify the data and make it more understandable, to increase the calculation speed, but also to increase the accuracy of the data mining result.
In our experiment, we found a phenomenon in both already symbolized ”tonnage” and “slew”, that is symbols are sometimes occurred continuously, and such phenomenon repeatedly occur. For example, in “tonnage”, symbol 1 is occurred in many intervals of 1 to 11031, 11041 to 29808, and so on.
We set 3 new attributes for the time series data, they are times, start position and end position. In this way, we reduce greatly the length of the time series. The new attribute of times represents the number of times that the symbol is occurred continuously, new attribute of start position means the starting time that the symbol is occurred continuously, new attribute of end position means the ending time that the symbol is occurred continuously.
Through MATLAB programming, we achieve dimensionality reduction for “tonnage” , and get a new dataset of 4 by 814 matrix. See table 6.5

cluster	1	0	1	2	.....	3	2
times	11033	1	1	5	.....	12	29
start position	1	11034	11035	11036	.....	86360	86372
end position	11033	11034	11035	11040	.....	86371	86400

Table 6.5: time series "tonnage" after dimensionality reduction

We use the same method to reduce the dimensionality of "slew", and get a new dataset of 4 by 316 matrix as shown in table 6.6

cluster	2	3	2	3	.....	2	1
times	28375	114	1381	88	.....	26	94
start position	1	28376	28490	29871	.....	86281	86307
end position	28375	28489	29870	29958	.....	86306	86400

Table 6.6: time series "slew" after dimensionality reduction

Table 6.5 and table 6.6 are two new datasets whose first lines demonstrate the cluster number to which the same symbol belongs, the 2nd lines of both tables mean the number of times that the symbol occurs in this interval, the 3rd lines of the tables are the start position which means the starting positions that the symbols start to occur in this interval, and the 4th lines mean the ending positions that the symbols stop occurring in this interval. We can see clearly that after dimensionality reduction these two new time series datasets "tonnage" and "slew" contain the same information that their original datasets have, but their sizes are greatly reduced, that means, we need much less storage space, less calculating time, and less cost to mine these two datasets.

6.4 Knowledge Discovery

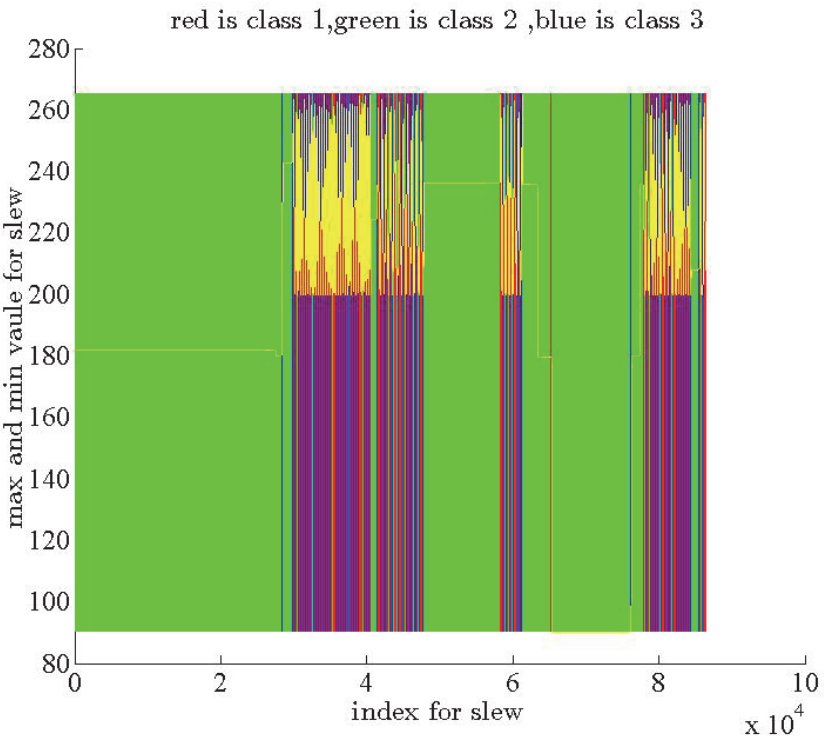

Figure 6.10: The patch diagram of "slew" after symbolization and dimensionality reduction

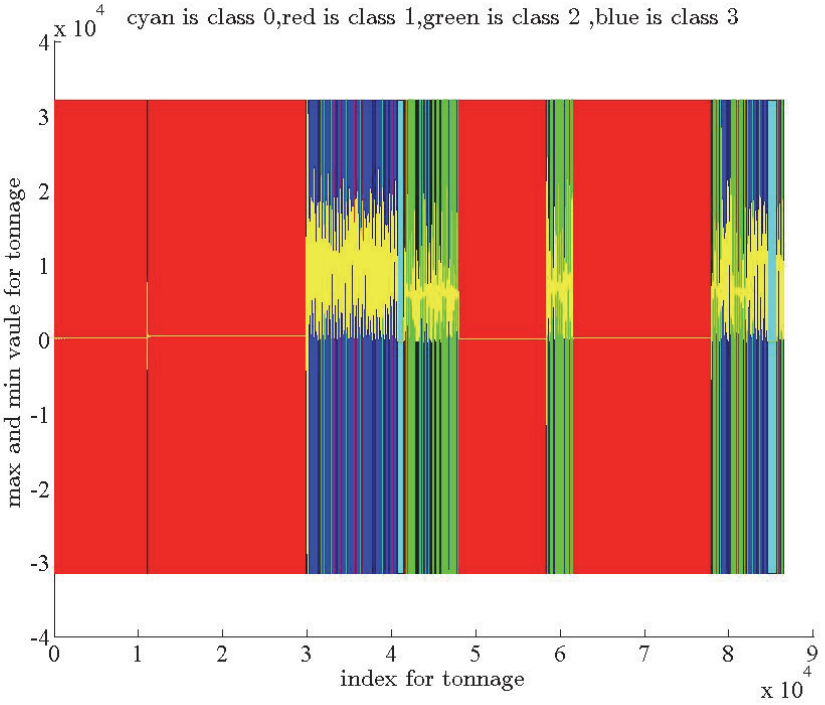

Figure 6.11: The patch diagram of "tonnage" after symbolization and dimensionality reduction

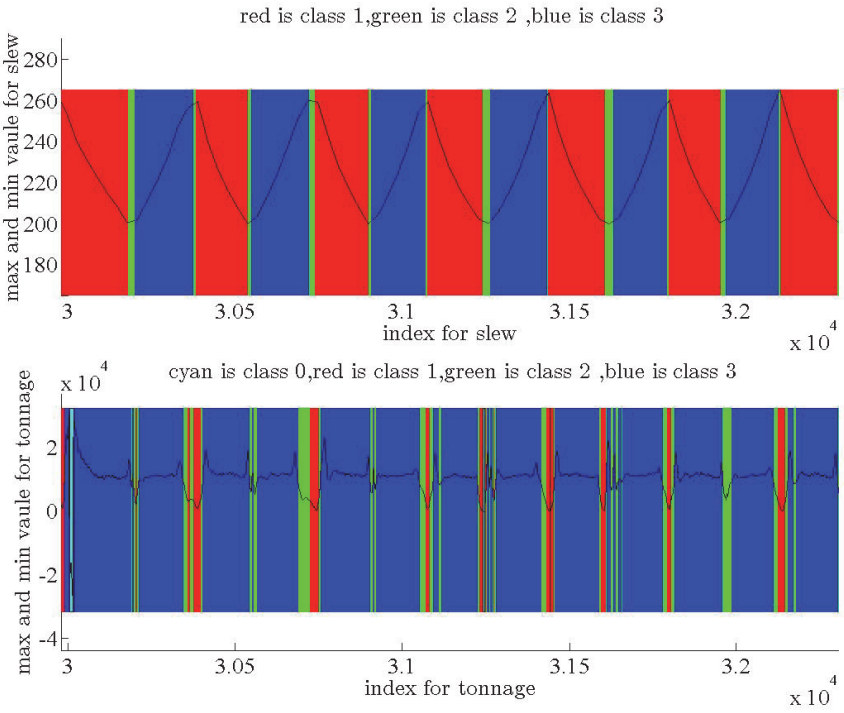

Figure 6.12: The patch diagram of "tonnage" and "slew" zoom in time

In Fig.6.10 and Fig.6.11, Fig.6.12,axis X represents attribute of time, the different clusters in "tonnage" and "slew" datasets are represented by different colors.

After symbolization there is a symbol at each time point of "tonnage" and "slew", but the same symbols 1 and 2 and 3 in "tonnage" and "slew" have different meanings. If we want to combine these two datasets to discover the knowledge, we should do the following work:

- Step 1

  We must re-define the symbols, that means re-defining clusters. We have used 4 symbols to represent "tonnage", and 3 symbols to represent "slew", that means now we need $3 \cdot 4 = 12$ symbols (clusters) to represent the two having been combined datasets "tonnage" and "slew".

		Cluster of 'slew'		
		1	2	3
Cluster of 'tonnage'	0	1	2	3
	1	4	5	6
	2	7	8	9
	3	10	11	12

Table 6.7: New clusters for "tonnage" and "slew"

In Tab.6.7 demonstrates at a time point, when the cluster of "tonnage" is 0, while the cluster of "slew" is 1, the new cluster is 1; When the cluster of "tonnage" is 0, while the cluster of "slew" is 2, then the new cluster is 2: And when the cluster of "tonnage" is 0, while the cluster of "slew" is 3, then the new cluster is 3, and so on. In this way, we get 12 new clusters, that means that we via MATLAB have obtained a new dataset of 2 by 86400, which represent all the information in time series "tonnage" and "slew", in another word, we have combined "tonnage" and "slew" into one time series dataset. See table 6.8

index	1	2	3	.....	86399	86400
cluster	5	5	5	.....	7	7

Table 6.8: the combination of symbolized representation of "tonnage" and "slew"

The first line of table 6.8 means the index of time, the 2nd line means the cluster that the time point belongs to.

- Step 2
  Now by the same dimensionality reduction as above mentioned we get again a new dataset of 4 by 1133 matrix. See table 6.9

cluster	5	2	5	8	.....	10	7
times	11033	1	1	5	.....	12	29
start position	1	11034	11035	11036	.....	86360	86372
end position	11033	11034	11035	11040	.....	86371	86400

Table 6.9: The combination of "tonnage" and "slew" after dimensionality reduction

- Step 3
  At last we do a mathematical statistics, by calculating out the number of occurring times of each cluster (class) in this new dataset or table 6.9, the mean value, the standard deviation, and the length or the time duration. These statistics information can all be obtained by MATLAB programming, as shown in Tab.6.10.

Cluster of "tonnage"	Cluster of "slew"	New cluster	frequ.	Mean	Standard deviation	Total [s]
0	1	1	16	4,87	1,69	78
0	2	2	21	116,52	97,61	2447
0	3	3	8	5,63	1,42	45
1	1	4	97	11,44	13,26	1110
1	2	5	124	460,32	1873,80	57080
1	3	6	92	10,61	15,57	976
2	1	7	145	37,42	61,36	5426
2	2	8	159	5,24	3,56	833
2	2	9	183	32,19	54,95	5890
3	1	10	107	59,49	54,04	6365
3	2	11	90	6,68	4,25	601
3	3	12	91	60,98	51,93	5549

Table 6.10: Conclusion to data mining of the two real datasets

6.5 Conclusion to Data Mining of Real Datasets

Tab.6.10 shows an example of data discovery via token statistics resulting from a merge of two symbol tables. We can see, cluster 5, a combination of cluster 1 in "tonnage" and cluster 2 in "slew", has the longest time durations of total length 57080s, that means in a whole day's production time the machine is mostly in static status. Then the total lengths of cluster 12, 10, 9, 7 are 5549s, 6365s, 5890s, and 5426s respectively. We can see these 4 clusters represent that the machine is moving with 2 different speeds to and fro. So the information we get from the mining of the two real time series datasets "tonnage" and "slew" collected from the monitoring system of real production activities is that, the most of time in a whole working day the machine is in static position, and the rest of the time in the day it moves to and fro with 2 different speeds.

Chapter 7

Conclusion

In this work, there is an overview of several commonly-used algorithms about data mining, which focus on application for Mechatronic systems. And also two real data sets recorded on a production machine are used to make an example, which describes a typical data mining process.
In this thesis, some basic concepts and knowledge about data and data mining has been presented, such as tasks of data mining, data definition, data attribute type, measurements of the data and so on. Since most of the data need some preprocessing before mining, in this thesis there are also some most important preprocessing methods introduced.
Time series data is everywhere in our daily life, and this topic is most important for mechatronic system, therefore, this thesis is also focused on time series data mining (TSDM). Different methods are suitable for different kinds of data, but generally speaking, time series data clustering methods are most useful and popular.
Finally after this overview, the thesis gives an example to show each step of data mining of two real data sets. In this data mining process, we use some suitable application for each step. For example, data cleaning is achieved by using the most probable value to fill in missing values, clustering is done with K-means and symbolisation is done with SAX (Symbolic Aggregate Approximation) and so on.
In the end of knowledge discovery of data mining, we can easily see the in-

formation about machine status hidden in the original data. In other words, we have transformed two sets of original, complicated and directly incomprehensive data into interesting, useful and understandable data. The example in this thesis shows, that through data mining we have obtained information about a whole day's working status of the machine, which was being observed, from two big original time series data sets.
The aim of this thesis is to give engineers a fast overview of solving big data problem in mechatronic systems.

Bibliography

[1] A.K.Jain, M.N.Murty, and P.J.Flynn. *Data clustering: a review.* ACM Computing Surveys(CSUR), 1999.

[2] John Aldrich and Miller Jeff. Earliest uses of symbols in probability and statistics.

[3] Christopher M Bishop. *Ratten Recognition and Machine Learning.* Springer, 2006.

[4] Li Chuan and Zhang Yong Hui. *Data Mining Practical Machine Learning Tools and Techniques.* China Machine Press, 2014.

[5] EOGH E. Exact indexing of dynamic time warping. In *Porc of 28th International Conference on Very Large Databases*, pages 406–417, 2002.

[6] EOGH E. Data mining and machine learning in time series database. In *Porc of the 5th Industrial Conference on Data Mining (ICDM)*, 2005.

[7] David A Freedam. *Statistical Models: Theory and Practice.* Cambridge University Press, 2009.

[8] SPIPADA S G, REITE R E, and HUNTER J. Segmenting time series for weather forecasting. In *Porc of ES.*, 2002.

[9] Randy Geobel, Joerg Siekmann, and Wolfgang Wahlster. *Advances in Knowledge Discovery and Data Mining.* Springer-Verlag Berlin Heidelberg, 2009.

[10] P.G Guest. *Numerical Methods of Curve Fitting*, chapter 7: Estimation of Polynomial Coefficients, page 147. Cambridge University Press, 2012.

[11] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan kaufmann, 2006.

[12] Jiawei Han and Micheline Kamber. *Data Mining*. Diane Cerra, 2006.

[13] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining Conceptes and Techniques*. Morgan Kaufmann, 2012.

[14] Liu Han, Liu Ding, and Li Qi. *Nonlinear prediction based on SVM time series*. Systems Engineering Theory and Pracitce, 2005.

[15] D.J. Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. MIT Press, 2001.

[16] YU J, HUNTE R J, and REITE R E. *Recognising visual pattems to communicate gas turbine time-series data*. Applicacations and Innovations in Interlligent Systems. London : Springer, 2002.

[17] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: areview. In *ACM Computing Surveys (CSUR)*, pages 264–323, 1999.

[18] E. Keogh and S. Kasetty. On the need for time series datamining benchmarks: a survey and empirical demonstration. In *DataMining and Knowledge Discovery*, pages 349–371, 2003.

[19] Mark Last, Abraham Kandel, and Horst Bunke. *Data Mining in Time Series Databases*. World Scientific, 2004.

[20] Luo Sheng Ling, Ma Jun, and Pan Li Ming. *Data Mining Theory and Technology*. Publishing House of Electronice Industry, 2013.

[21] Oded Maimon and Lior Rokach. *The Data Mining and Knowledge Discovery Handbook*. Springer Science Business Media,inc, 2005.

[22] Mohd.Shahnaeaz, Ashish Ranjan, and Mohd Danish. Temporal data mining : An overview. In *International Journal of Engineering and Advanced Technology(IJEAT)*, 2011.

[23] Paul O'Leary and Matthew Harker. Discrete orthogonal polynomial tool-box: Dopbox version 1.8. `http://www.mathworks.com/matlabcentral/fileexchange/`. Online; 11 Apr 2013 (Updated 04 Mar 2014).

[24] Paul O'Leary and Matthew Harker. localconv: Use conv to implement local operators with correct end poins. version 1.0. `http://www.harkeroleary.org`. Online; 17.Sept 2014.

[25] Witold Pedrycz and Shyi-Ming Chen. *Time Series Analysis, Modeling and Applications: A Computational Intelligence Perspective*. Springer Heidelberg New York Dordrecht, 2012.

[26] AGRAWAL R, FALOUTSOS C, and SWAMI A. Efficient simlarity search in sequence databases. In *Porc of the 4th International Conference on Foundations of Data Organization and Algorithms*, pages 69–84, 1993.

[27] Savitzky and Abraham. A historic collaboration. *Analytical Chemistry*, 61 (15), 1989.

[28] A. Savitzky and M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36 (8), 1964.

[29] Piatesky Shapiro and Gregory. *Knowledge Discovery in Databases*. AAA/MIT Press, 1991.

[30] Jean Steinier, Yves Termonia, and Jules Deltour. Smoothing and differentiation of data by simplified least square procedure. *Analytical Chemistry*, 44 (11), 1972.

[31] E.T Whittaker and G Robinson. *The Calculus Of Observations*. Blackie and Son, 1924.

[32] HU Xiao-lin and CHEN xiao yun. Frequent subsequence mining in the time series based on symbolic representation. *Computer Engineering*, 34(10), 2008.

[33] HUHTALA Y, K RKK INEN J, and TOIVONEN H. Mining for similarities in aligned time series using wavelets. In *Porc of Data Mining and Knowledge Discovery: Theory, Tools and Technology*, pages 150–160, 1999.

[34] Chen zhe, Feng tianjing, and Zhang haiyan. *Time series analysis and phase space reconstruction*. Computer Research and Development, 2001.

[35] Jan Zytkow and Jan Rauch. *Principles of Data Mining and Knowledge Discovery*. Springer-Verlag Berlin Heidelberg, 1999.

Appendix

Preprocessing

```
clear all;
close all;
load slewTonnage
%
tonnage1=reshape(tonnage,1,length(tonnage));
time1=1:length(tonnage);
rawdata_tonnage=[tonnage1;time1];
%
%
Inds=find(isnan(tonnage1));
allInds=1:length(tonnage);
useInds=setdiff(allInds,Inds);
%
%
tonnagewithoutNaN=tonnage1(useInds);
timewithoutNaN=time1(useInds);
rawdatawithoutNaN_tonnage=[tonnagewithoutNaN;timewithoutNaN];
%
n=length(Inds);
a=linspace(tonnage(Inds(1)-1),tonnage(Inds(n)+1),n);
```

```
tonnage3=tonnage1;
for i=1:length(Inds)
    tonnage3(Inds(i))=a(i);
end
rawdataafterPreprocess_tonnage=[tonnage3;time1];
save rawdataafterPreprocess_tonnage;
save rawdata_tonnage rawdata_tonnage;
save NaNpostion_for_tonnage Inds;
```

Clusting

```
%% Clear up the workspace
%
close all;
clear all;
%
%% Load the Data
%
%load slewTonnage;
load rawdataafterPreprocess_tonnage
%
tonnage=rawdataafterPreprocess_tonnage(1,:);
symbols=zeros(size(tonnage));
%
minT = min( tonnage );
maxT = max( tonnage );
%
nrBins = 51;
bins = linspace( 0, maxT, nrBins );
%
cts = hist( tonnage, bins );
```

```
%
%% Plot the histogram
%
fig1 = figure;
A(1) = subplot(2,1,1);
H = bar( bins, cts, 'r');
%hist(tonnage,bins);
xlabel('Tonnage');
ylabel('frequency');
%
range = axis;
top = 6e3;
axis( [range(1:2), 0, top] );
%
%% Get some graphical inputs
%
nrPts = 2;
[x, y] = ginput(nrPts);
[x, inds] = sort( x );
y = y(inds);
%
%% the the selected pointd
hold on;
range = axis;
for k=1:nrPts
    plot( [x(k), x(k)], range(3:4), 'k');
end;
%
%% Segment the data and compute teh stats
level = [0; x; maxT];
%
for k=1:(length( level )-1);
    inds = find( (tonnage >= level(k) ) & (tonnage < level(k+1)) )
```

```
    m = mean( tonnage(inds) );
    s = std( tonnage(inds) );
    means(k) = m;
    sigmas(k) = s;
    cut(k,1) = level(k);
    cut(k,2) = level(k+1);
    symbols(inds)=k;
end;
xt=length(level);
inds1=find(tonnage==maxT);
symbols(inds1)=xt-1;
means(1) = 0;
%
%
%% Producte the plots for the normal distributions
color(1) = 'r';
color(2) = 'g';
color(3) = 'b';
%
A(2) = subplot(2,1,2);
for k=1:length(means)
    y = exp( -((bins - means(k)) / sigmas(k)).^2);
    plot( bins, y, color(k));
    hold on;
end;
linkaxes( A , 'x');
axis tight;
xlabel('Tonnage');
ylabel('probability');
%nCts = normalize( cts );
time=1:length(symbols);
tonnage_with_class=[tonnage;time;symbols];
save tonnage_with_class tonnage_with_class
```

Dimensionality Reduction

```
close all;
clear all;
load tonnage_with_class;
tonnage=tonnage_with_class(1,:);
tonnage_class=tonnage_with_class(3,:);
n=length(tonnage_class);
class=[tonnage_class(1)];
times=[];
time=1;
endposition=[];
startposition=[];
for i=1:n-1
    if (tonnage_class(i)==tonnage_class(i+1))
        time=time+1;
    else
        startposition=[startposition,(i-time+1)];
        endposition=[endposition,i];
        class=[class,tonnage_class(i+1)];
        times=[times,time];
        time=1;

    end
end
if (tonnage_class(n)==tonnage_class(n-1))
    t=length(times);
    times(t+1)=time;
    startposition(t+1)=n-time+1;
    endposition(t+1)=n;
else
```

```
    t=length(times)
    times(t+1)=1;
    startposition(t+1)=n;
    endposition(t+1)=n;
end
tonnage_classshort_with_time_start_end= ...
    [class;times;startposition;endposition];
save tonnage_classshort_with_time_start_end ...
     tonnage_classshort_with_time_start_end
```

Knowledge Discovery

```
close all;
clear all;
load dslew_classshort_with_time_start_end;
load tonnage_classshort_with_time_start_end;
load rawdataafterPreprocess_slew;
load rawdataafterPreprocess_tonnage;
load dslew_with_class
ds_start=dslew_classshort_with_time_start_end(3,:);
ds_end=dslew_classshort_with_time_start_end(4,:);
s=rawdataafterPreprocess_slew(1,:);
ds_class=dslew_classshort_with_time_start_end(1,:);
tonnage_start=tonnage_classshort_with_time_start_end(3,:);
tonnage_end=tonnage_classshort_with_time_start_end(4,:);
tonnage=rawdataafterPreprocess_tonnage(1,:);
tonnage_class=tonnage_classshort_with_time_start_end(1,:);

fig1=figure;
A(1) = subplot(2,1,1);
ymax=max(s);
```

```
ymin=min(s);

for i=1:length(ds_start-1);
    x=[];
    y=[];
    x=[x;ds_start(i);ds_start(i);ds_end(i)+1;ds_end(i)+1];
    %mintemp=min(ds(ds_start(i):ds_end(i)));
    %maxtemp=max(ds(ds_start(i):ds_end(i)));
    %y=[y;mintemp;maxtemp;maxtemp;mintemp;];
    y=[y;ymin;ymax;ymax;ymin;];
    if ds_class(i)==1
        patch(x,y,'r','edgecolor','none');
        hold on
    elseif ds_class(i)==2
        patch(x,y,'g','edgecolor','none');
        hold on
    elseif ds_class(i)==3
        patch(x,y,'b','edgecolor','none');
        hold on
    end
end

%last one block
i=length(ds_start);
x=[ds_start(i);ds_start(i);ds_end(i);ds_end(i)];
y=[ymin;ymax;ymax;ymin;];
if ds_class(i)==1
        patch(x,y,'r','edgecolor','none');
        hold on
    elseif ds_class(i)==2
        patch(x,y,'g','edgecolor','none');
        hold on
    elseif ds_class(i)==3
```

```
        patch(x,y,'b','edgecolor','none');
        hold on
end
plot(s,'y');
title('red is class 1,green is class 2 ,blue is class 3') ;
xlabel('index for slew');
ylabel('max and min vaule for slew');
%%
%fig2=figure;
A(2) = subplot(2,1,2);
ytmax=max(tonnage);
ytmin=min(tonnage);
for i=1:length(tonnage_start-1);
    xt=[];
    yt=[];
    xt=[xt;tonnage_start(i);tonnage_start(i);...
        tonnage_end(i)+1;tonnage_end(i)+1];
    yt=[yt;ytmin;ytmax;ytmax;ytmin;];
    if tonnage_class(i)==0
        patch(xt,yt,'c');
        hold on
    elseif tonnage_class(i)==1
        patch(xt,yt,'r','edgecolor','none');
        hold on
    elseif tonnage_class(i)==2
        patch(xt,yt,'g','edgecolor','none');
        hold on
    elseif tonnage_class(i)==3
        patch(xt,yt,'b','edgecolor','none');
        hold on
    end
end
%last one block
```

```
i=length(tonnage_start);
xt=[tonnage_start(i);tonnage_start(i);...
    tonnage_end(i);tonnage_end(i)];
yt=[ymin;ymax;ymax;ymin;];

if tonnage_class(i)==0
        patch(xt,yt,'c');
        hold on
    elseif tonnage_class(i)==1
        patch(xt,yt,'r','edgecolor','none');
        hold on
    elseif tonnage_class(i)==2
        patch(xt,yt,'g','edgecolor','none');
        hold on
    elseif tonnage_class(i)==3
        patch(xt,yt,'b','edgecolor','none');
        hold on
end
plot(tonnage,'y');
title('cyan is class 0,red is class 1,...
      green is class 2 ,blue is class 3') ;
xlabel('index for tonnage');
ylabel('max and min vaule for tonnage');
linkaxes( A, 'x');
```

Mathematical statistics

```
close all;
clear all;
load slewandtonnage_class_start_end;
class=slewandtonnage_class_start_end(1,:);
```

```
length_eachclass=slewandtonnage_class_start_end(2,:);
maxClass=max(class);
class1=(1:maxClass)';
no=[];
for i=1:maxClass
    indx=find(class==i);
    n=length(indx);
    no=[no;n];
end
table=[class1';no']';
length_class=[];
for i=1:maxClass
   indx1=find(class==i);
   n=length(indx1);
   for j=1:n
       length_class(i,j)=length_eachclass(indx1(j));
   end
end
table1=table;
sumlength=sum(length_class,2);
for i=1:maxClass
    mean(i)=sumlength(i)/no(i);
end
mean=mean';
table2=[table1 mean];
s_length=[];
std_length=std(length_class,0,2);
table3=[table2 std_length];
statistics_table=[table3 sumlength];
save statistics_table statistics_table
```