Research on pattern recognition method based on the analysis of large big quality

Huang Yanmei, Wu Qinlan

School of Internet of Things Engineering, Jiangxi College Of Engineering, JiangXi, 338029, China

Corresponding author's e-mail: huangyanmeim@126.com

Received 10 September 2013, www.cmnt.lv

Abstract

As the development of Internet, mobile Internet and networking, we have effectively ushered in an era of mass data. Analysis of research firm IDC released a new digital study reports, this report shows that total global information every two years, will grow 1 time. Therefore, as data growing, how to manage huge amounts of data and analyses has become a very important and urgent needs. Data quality is the basis for conclusion validity and accuracy of the data analysis is the most important prerequisite and guarantee. Pattern recognition development in the 1960s in in signal processing, artificial intelligence, Cybernetics, computer science and other disciplines with its high speed, high accuracy, and high efficiency characteristics of large data processing has its unique advantages.

Keywords: big data, pattern recognition; internet

1 Introduction

Big Data[1] is another disruptive technology revolution in the IT industry after cloud computing, Internet of things . it will have an enormous impact on the state governance, corporate decisions, personal life style, organizational and business processes and so on. Next, introduce the meaning and revolution of the big data

1.1 DEFINITION OF BIG DATA

"Big data", research firm Gartner gives this definition. "Big data" is the information assets which need for new processing modes to have more decision-making power, insight into the ability to detect and process optimization of mass, high rates of growth and diversification.

The 4 typical features of the data, the so-called 4 "v", that is, variety, volume, velocity and value. Variety means, the data type, you should include structural and non-structural data, aggregate volume refers to for analysis must be very large amounts of data and velocity is the speed of data processing must be quickly, value reflected in the value of low density and high commercial value.

1.2 THE CHANGES CAUSED BY LARGE DATA

Big data gives us the ability to predict the future through data analysis. Data analysis is good for the national development plan, the enterprise to understand customer needs and grasp market trends. Data analysis, beginning with data quality, data quality includes data completeness, consistency, accuracy, timeliness. Data quality analysis involves three aspects: data collection, processing and application.

The 5 levels of data analysis are indicated as: First, Visual analysis, visualize the data and let the data speak; the second is data mining algorithms, data for the machine; Three, semantic engine, parse, extract, analyze unstructured data; Four is predictive analysis; Five is data quality management [2].

The big data in the information age, big data supporting the development of national security and an important strategic resources. Who occupy the advantages of information technology, which will maximize the storage, mining and use of "big data", firmly grasp the "big data" development and utilization initiative? It was vividly called "data sovereignty".

For our country on big data, embodied have distributed in four aspects: first, through the implementation of "through several management" can improve decision-making capabilities. For example, by analyzing the mobile user roaming during the Spring Festival, master scale of population movements and migration rules, it helps make decision for traffic management, rail transport, public safety, management and other decision makers. Second is based on big data applications, gradually open up public data, to create a transparent Government, improve the credibility of the Government. Third, based on data analysis, monitoring major social events, build a scientific early warning surveillance system to good for people's livelihood. Use of data throughout the Government and in all areas of society, in health care, food safety, road traffic, geological hazards, social opinion, information security, homeland security and other areas of intelligence analysis, effective implementation is a major security risk, crises, prevention and early warning. Four industries all need to strengthen the sense of big data, using data to improve efficiency, enhance the level of refinement and intelligence to advance from made in China to create in China, and creating greater value.

2 Definition of pattern recognition

A. Introduction of pattern recognition

Pattern means that things exist in time and space which you can observe and with temporal or spatial distribution. Pattern recognition means that using computer implements

the analysis, description, and judgment, recognition of the objects or phenomena by human [3]. Pattern recognition refers to various forms of representation of objects or phenomena (numeric, text, and logical relationship) information processing and analysis, and it's the process of description, identification, classification, and interpretation of objects or phenomena, and it is an important part of the information sciences and artificial intelligence.

Pattern recognition is a field within the area of machine learning. Alternatively, it can be defined as "the act of taking in raw data and taking an action based on the category of the data" . As such, it is a collection of methods for supervised learning.Pattern recognition aims to classify data (patterns) based on either a priori knowledge or on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations, defining points in an appropriate multidimensional space.

A complete pattern recognition system consists of a sensor that gathers the observations to be classified or described; a feature extraction mechanism that computes numeric or symbolic information from the observations; and a classification or description scheme that does the actual job of classifying or describing observations, relying on the extracted features.

The classification or description scheme is usually based on the availability of a set of patterns that have already been classified or described. This set of patterns is termed the training set and the resulting learning strategy is characterized as supervised learning. Learning can also be unsupervised, in the sense that the system is not given an a priori labelling of patterns, instead it establishes the classes itself based on the statistical regularities of the patterns.

The classification or description scheme usually uses one of the following approaches: statistical (or decision theoretic), syntactic (or structural). Statistical pattern recognition is based on statistical characterizations of patterns, assuming that the patterns are generated by a probabilistic system. Structural pattern recognition is based on the structural interrelationships of features.

Typical applications are automatic speech recognition, classification of text into several categories (e.g. spam/non-spam email messages), the automatic recognition of handwritten postal codes on postal envelopes, or the automatic recognition of images of human faces. The last three examples form the subtopic image analysis of pattern recognition that deals with digital images as input to pattern recognition systems.

Pattern recognition has relations with statistics, psychology, Linguistics, computer science, biology, Cybernetics. And also have Intersect relations with the study of artificial intelligence and image processing [4-5]. Such as adaptive or self-organizing pattern recognition systems include artificial intelligence learning mechanism, artificial intelligence research scene understanding, natural language understanding also includes pattern recognition problem. Another example, preprocessing and feature extracting part of the pattern recognition using image processing techniques; image analysis and pattern recognition can also be used in image processing technology [6]. Figure 1 shows the pattern recognition system training process.

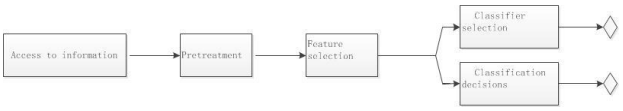

FIGURE 1 Pattern recognition system training process

TABLE 1 Pattern recognition system training process.

samples	x_1	x_2	…	x_n
X_1	X_{11}	X_{12}	…	X_{1n}
X_2	X_{21}	X_{22}	…	X_{2n}
…	…	…	…	…
X_n	X_{n1}	X_{n2}	…	X_{nn}

B. Pattern sample representation method

(1) Vector representation: suppose a vector with n variable,

$$X = (X_1, X_2, ..., X_n)^T$$

(2) Matrix representation: N samples and n variables, As shown in table 2.1

(3)Geometry representations: Indicated by figure 2.

One dimensional:

$X_1 = 1.5$, $X_2 = 3$

Two dimensional:

$$X_1 = (x_1, x_2)^T = (1,2)^T$$

$$X_2 = (x_1, x_2)^T = (2,1)^T$$

Three dimensional:

$$X_1 = (x_1, x_2, x_3)^T = (1,1,0)^T$$

$$X_2 = (x_1, x_2, x_3)^T = (1,0,1)^T$$

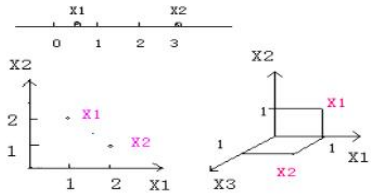

FIGURE 2 Pattern recognition system training process

3 Big data mining based on Pattern recognition

A. Data mining and application

Data mining [7] (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

In many data stream mining applications, the goal is to predict the class or value of new instances in the data stream given some knowledge about the class membership or values of previous instances in the data stream. Machine learning techniques can be used to learn this prediction task from labeled examples in an automated fashion. Often, concepts from the field of incremental learning, a generalization of Incremental heuristic search are applied to cope with structural changes, on-line learning and real-time demands. In many applications, especially operating within non-stationary environments, the distribution underlying the instances or the rules underlying their labeling may change over time, i.e. the goal of the prediction, the class to be predicted or the target value to be predicted, may change over time. This problem is referred to as concept drift.

Data mining is needed first collected a large amount of data in a business environment, and requires knowledge of mining is valuable [8]. Value to the business, nothing more than three conditions: lower costs; increase revenue; to increase the share price.

But how can we find the most value of data to us in the huge data? Fortunately, pattern recognition can help us achieve our goals. First, we need to sample and then classification the data, and analyze the data.

B. Similarity and classification

1. The similarity between two samples have to meet the following requirements [9].
   (1) Non-negative
   (2) Similarity between the two samples should be the maximum
   (3) Measurements should satisfy the symmetry
   (4) Similarity should be the monotone functions of distance between points
2. Using kinds of distance to express similarity
   Given two samples:

$$X_i = (x_{i1}, x_{i2}, ..., x_{in})^T$$

$$X_j = (x_{j1}, x_{j2,...,} x_{jn})^T$$

Distance of absolute value:

$$D_{ij} = \sum_{k=1}^{n} | X_{ik} - X_{jk} |$$

Euclidean distance:

$$D_{ij} = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2}$$

3. Mingkaofusiji distance:

$$D_{ij}(q) = \left(\sum_{k=1}^{n} | X_{ik} - X_{jk} |^q \right)^{1/q}$$

Where q=1,this is distance of absolute value and q=2, is the Euclidean distance.

4. Chebyshev distance:

$$D_{ij}(\infty) = \max_{1 \le k \le n} | X_{ik} - X_{jk} |$$

5. Mahalanobis distance:

$$D_{ij}(M) = \sqrt{(X_i - X_j)^T / \sum^{-1} (X_i - X_j)}$$

6. Cosine: It represents a small angle between the samples as a class.

$$C_{ij} = \frac{\sum_{k=1}^{n} X_{ik} X_{jk}}{\sqrt{(\sum_{k=1}^{n} X_{ik}{}^2)(\sum_{k=1}^{n} X_{jk}{}^2)}}$$

7. Correlation coefficient: we have to standardization the data before seeking the correlation coefficient.

$$r_{ij} = \frac{\sum_{k=1}^{n} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{\sqrt{\sum_{k=1}^{n} (X_{ki} - \bar{X}_i)^2 \sum_{k=1}^{n} (X_{kj} - \bar{X}_j)^2}}$$

C. Data analysis processing method – pattern recognition

When we combine the data after the mining category, then we needs to use pattern recognition to identify a data quality is good or bad. Pattern recognition refers to various forms of representation of objects or phenomena (numeric, text, and logical relationship) information processing and analysis, and it is the process of description, identification, classification, and interpretation of objects or phenomena.

Commonly used method is decision-making theory and syntactic methods. Decision-making theory [10]: first will be digital the identified objects, and transforms to the digital information which can apply in the computer system. A mode usually requires a large amount of information to express. Figure 3.1 shows the processing.

Syntactic methods [11]: Calculated by the eigenvector corresponding to different categories of discriminant function values, entity classification by discriminant function values by syntactic methods. Syntax is also known as structure or Linguistics methods, The basic idea is to put a pattern description as a combination of simpler sub-patterns, sub-pattern can be described as a combination of simpler sub-patterns, and eventually got a tree structure description, at the bottom of the most simple mode called mode primitive. Selected in the syntactic primitives of the problem is equivalent to select the feature in the decision theory approach problems. Figure 3 and 4 shows the processing.

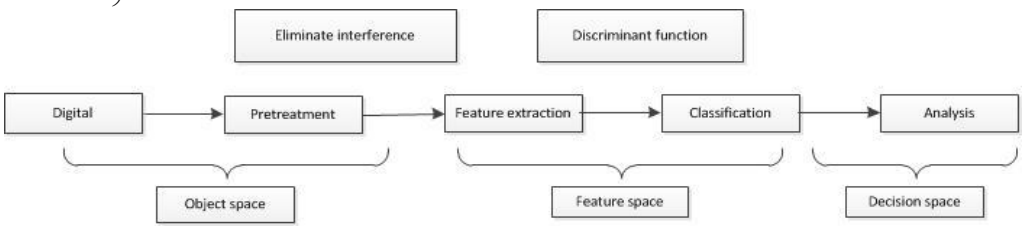

FIGURE 3 The basic progressing

FIGURE 4 The basic progressing

4 Conclusions

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications.

The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on.

As development of our society, big data will be more and more important in our economic. So strengthen the analysis of big data and research it, it will be better for our country to make right economic decision. Pattern recognition is a good method to analysis big data. If we can make full use of it, we will be able to benefit from it.

References

[1] Big data<http://en.wikipedia.org/wiki/Big_data>.
[2] Han, MichelineKamber 2001 DataMining: Conceptsand Techniques[M] *USA: Morgan Kaufmann Publishers*. 24-32
[3] David W Hosmer 2000 Applied logistic regression *USA Wiley2 Interscience Publication* 124-31
[4] Janardhana Iyengar R S;Sastry V V 1995 Fuzzy logic based soft - start for induction motor drives Industry Applications *Thirtieth IAS Annual Meeting, IAS'95 Conference Record* 192-9
[5] Jain A K, Duin R P W, Mao J 2000 Statistical pattern recognition: a review *IEEE Trans. PAMI* **22(**1) 4-37
[6] Jones V M J 2001 Rapid object detection using a boosted cascade of simple features *Proc. CVPR, Hawaii*.**1** 511-8
[7] Liu G X 2003 Aggregate homology methods for solving sequential max-min problems, complementarity problems and vibrational inequalities *PhD thesis Jilin university of China* 61-3
[8] Bins J, Draper BA 2001 Feature selection from huge feature sets, *Proc. 8th ICCV* **2** 159-65
[9] Dai, Yuen P C 2003 Regularized discriminant analysis and its application to face recognition *Pattern Recognition* **36(**3) 845-7
[10] Yu J, Yang 2001 A direct LDA algorithm for high-dimensional data--with application to face recognition *Pattern Recognition* **34(**10) 2067-70
[11] Sarawagi S, Agrawal R, Megiddo N 1998 Discovery-Driven Exploration of OLAP Data Cubes 12-22

Authors

Wu Qinlan, born in May 1982, Xinyu City, Jiangxi Prov, China

Current position, grades: Lecturer
University studies: Computer science
Scientific interest: Cloud computing
Experience: From 2006 to 2014, working in Jiangxi college of engineering.

Huang Yanmei, born in January 1979, Nanchang City, Jiangxi Prov, China

Current position, grades: Lecturer
University studies: Computer science
Scientific interest: Cloud computing
Experience: From 2002 to 2014, working in the computer related