J. Chosun Natural Sci.
Vol. 7, No. 1 (2014) pp. 57 – 61
http://dx.doi.org/10.13160/ricns.2014.7.1.57

A Study on Efficient Cluster Analysis of Bio-Data Using MapReduce Framework

Sowol Yoo, Kwangok Lee, and Sanghyun Bae†

Abstract

This study measured the stream data from the several sensors, and stores the database in MapReduce framework environment, and it aims to design system with the small performance and cluster analysis error rate through the KM-SVM algorithm. Through the KM-SVM algorithm, the cluster analysis effective data was used for U-health system. In the results of experiment by using 2003 data sets obtained from 52 test subjects, the k-NN algorithm showed 79.29% cluster analysis accuracy, K-means algorithm showed 87.15 cluster analysis accuracy, and SVM algorithm showed 83.72%, KM-SVM showed 90.72%. As a result, the process speed and cluster analysis effective ratio of KM-SVM algorithm was better.

Key words: MapReduce, k-NN, K-means, SVM, KM-SVM, Cluster Analysis

1. Introduction

As the development and ubiquitous environment establishment of the current sensor network technology are commercialized, the efforts to obtain the useful added value information by analysis of the stream data are conducting actively recently. As the representative stream data process system, there are STREAM, Aurora for the network traffic monitoring, Gigascope etc. and in addition, the various stream process systems are developed[1].

The distribution of stream data may be changed by the big change of data or the pattern which is not occurred frequently as time goes by, and it has usually non-linear distribution in real-world. In addition, the stream data has the ordered characteristic, so it can be regarded as the time series data. The time series data can obtain the useful information through the analysis. The time series data analysis can be conducted through the existing data. Likewise, the useful information can be extracted through the decision making by the analysis on the stream data. In addition, the data mining to find the useful information in the high-risk patients' vital signs data effectively in the process became important field.

The data mining provides the necessary service to the patients by analysis on data by the high-risk patient's condition, so the management of high-risk patients through the prediction will be useful. It is known that the K-means algorithm is efficient for the cluster analysis on this dat. K-means algorithm compares the distance with the average value, so the improvement of the accuracy of the cluster analysis can be seen. But this method needs many data, but there is problem that is difficult to predict when learning is completed, and the performance speed decreases. In order to complement this problem, the SVM (Support Vector Machine) based on the theory of the statistical learning is the algorithm to accomplish the classification analysis and estimation, it is very functional and it is recognized as the efficient technique[2].

Therefore, this study measured the stream data from the several sensors, and stores the database in MapReduce framework environment, and it aims to design system with the small performance and cluster analysis error rate through the KM-SVM algorithm.

This study was configured as follows. Chapter 2 shows the configuration and design of system, and Chapter 3 shows the performance assessment and

Department of Computer Science & Statistics, Chosun University, Gwangju

†Corresponding author : shbae@chosun.ac.kr
(Received : February 18, 2014, Revised : March 13, 2014, Accepted : March 25, 2014)

experiment results. Chapter 4 suggested the conclusion and future direction of research.

2. Configuration and Design of System

This study aims to classify the vital data of the diabetic patient by using KM-SVM algorithm after obtaining the vital data by using a large number of sensor(blood pressure, pulse, body temperature, blood sugar) for the stream vital data.

Fig. 1 shows the configuration of system which classifies the vital data by the parallel processing using the MapReduce function and obtained vital data through the sensor.

2.1. Mapreduce Management System

This study shows the parallel processing system using the MapReduce framework. The mapper and the reducer are the interface as the core of the MapReduce programing environment. The mapper takes role to mapping input key and the value as the middle key value pair. The map task takes role to change the input record into the middle record. The converted middle record has not to be the same type with the input record. The given input pair may has nothing or may has a lot of vital data of multiple output pair as the value of the output pair[3].

Fig. 2. Shows the accomplishment flow of the MapReduce data process.

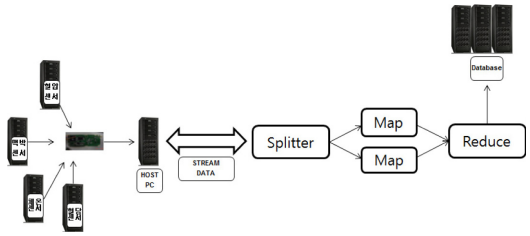

Fig. 1. System configuration.

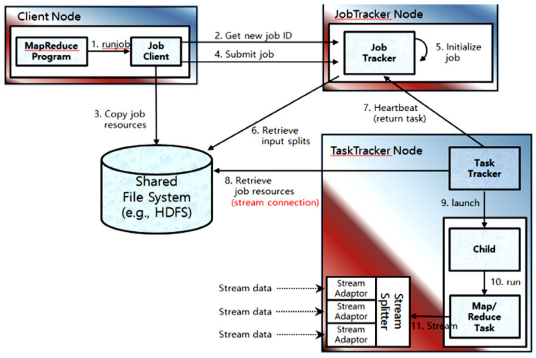

Fig. 2. The performance of MapReduce job flow.

2.2. Proposed KM-SVM Algorithm

To match the vital data with the diabetic patients' vital stream data cluster analysis, there is existing various learning algorithm, and the blood pressure, pulse, body temperature, blood sugar data were used for this experiment, it is composed with the non-linear data structure, the KM-SVM algorithm was used to solve the non-linear determining the problem with the multilayer perceptron structure.

The learning theory based KM-SVM shows the excellent performance of the cluster analysis and pattern recognition field. It takes role to classify whether the given data is applied to the specific category as the dual KM-SVM classification. In addition, it shows always constant excellent performance differently with the neural network classification system. In addition, it shows always constant excellent performance differently with the neural network classification system which has performance to be changed by learning with better extendability than the existing linear classification system in data process. Two sets can be defined as Fig. 3 in order to define the model for the classification and cluster.

KM-SVM is not for minimizing the sample error or finding simply classification flat, and there is high probability of proper classification and cluster about new data by maximizing the classification margin but the learning data.

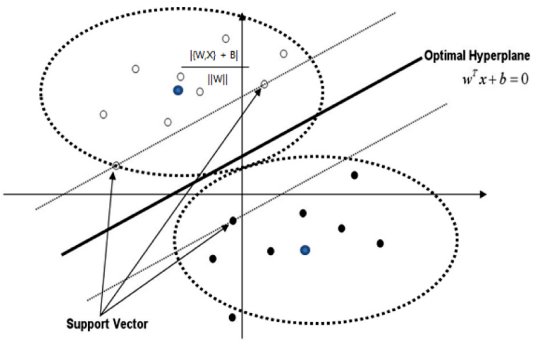

Fig. 3. KM-SVM Support vector.

Algorithm : KM-SVM
Number of data for comparative history learning : N
Inputs: Database including n object, k : Number of cluster
k_{i1} : blood pressure, k_{i2} : pulse, k_{i3} : body temperature, k_{i4} : blood sugar
Output: k cluster
Steps
(1) Select the number of cluster, k.
(2) Select the parameter r, C.
(3) Summary the data and obtain the cluster center by using clustering.
(4) Establish SVM through the center of cluster.
(5) (2) ~ (4) Select the optimal KM-SVM by the repetition of the process.

Fig. 4. KM-SVM Algorithm configuration.

This study configured the KM-SVM algorithm as Fig. 4.

3. Performance Assessment and Experimental Results

The experiment in this study is about the cluster measure of data mining using the MapReduce frame-work. The processor board used for the experiment is the Telos platform series, the experiment is conducted by using MSP430 MCU and CC24240 Radio Chip. By using total 10 nodes of 1 Sink node and 9 middle nodes, the transport process of computed data in the stream data storing unit can be conducted about the blood pressure, pulse, body temperature, blood sugar value once every 5 seconds. The collected data can be stored in database through the KM-SVM algorithm classification after query in FILE SYSTEM. The stored data was used for the cluster analysis through the k-NN, K-means, SVM, KM-SVM algorithm by the data mining. For the test of the implemented system, the four types of vital data such as blood pressures, pulse, body temperature, blood sugar were used among 2003 collected data. In this experiment, the diabetic patients were adopted as 1, the patients without diabetes were adopted as 0, the error rate of cluster analysis on blood pressure difference, pulse, body temperature were measured through the k-NN, K-means, SVM, KM-SVM algorithm.

Fig. 5 shows the variable form for the cluster analysis on the diabetes state as the dependent variable by the

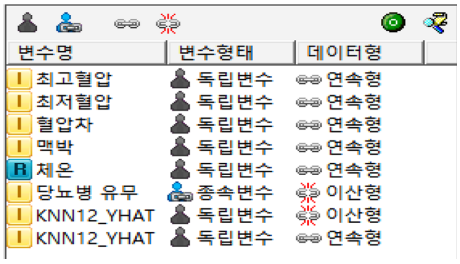

Fig. 5. Variables Biometric data.

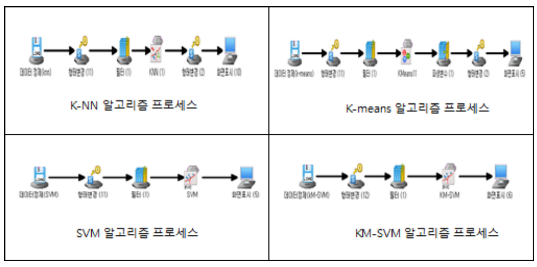

Fig. 6. Process of cluster analysis algorithm.

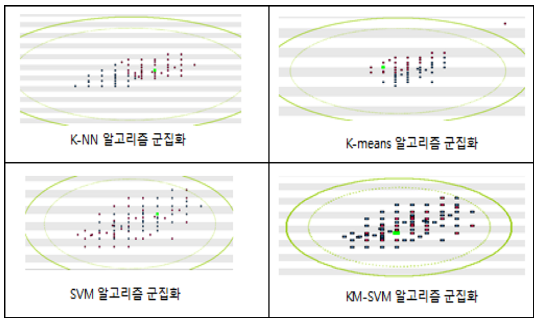

Fig. 7. Clustering of cluster analysis algorithm 1.

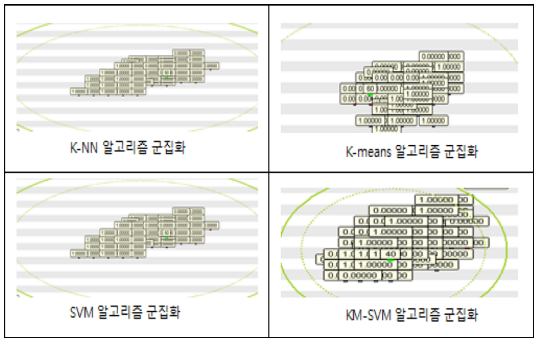

Fig. 8. Clustering of cluster analysis algorithm 2.

independent variable, and the maximal blood pressure, minimum blood pressure, blood pressure difference, pulse, body temperature were designated as the independent variable and data type continuously, and the diabetes state designated as the dependent variable.

Fig. 6 shows the process of each algorithm. The cluster analysis was experimented through the k-NN, K-means, SVM, KM-SVM algorithm.

Fig. 7 and Fig. 8 shows the clustering of maximal blood pressure, minimum blood pressure, pulse, body temperature of each algorithm. The data 1 means the cluster of diabetic patient, and data cluster set as 0 is not for the diabetic patient.

As the experiment results, the cluster analysis on KM-SVM, K-means, SVM, k-NN can be conducted. In the experiment results, it shows the good cluster accuracy with 90.72% KM-SVM algorithm. Table 1 and Fig. 9 show the accuracy of the cluster analysis algorithm.

Table 1. Accuracy of cluster analysis algorithm

Algorithm	Mis-classification	Accuracy
k-NN	20.71%	79.29%
K-means	12.85%	87.15%
SVM	16.28%	83.72%
KM-SVM	9.28%	90.72%

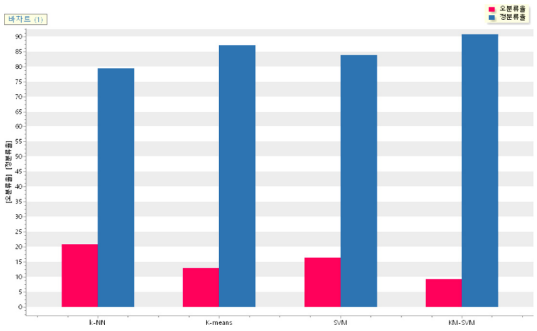

Fig. 9. Accuracy graph of cluster analysis.

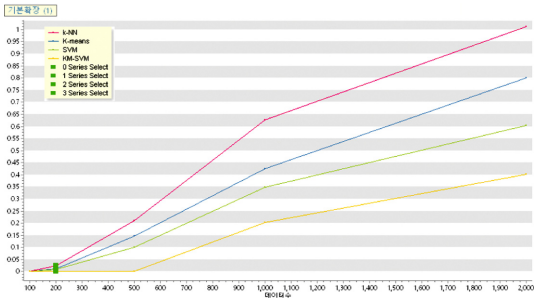

Fig. 10. Performance of the algorithm.

Fig. 10 shows the experiment results graph showing the performance assessment of algorithm. It is the experiment results graph to process many data per second in the order of KM-SVM, SVM, K-means, k-NN algorithm. In the graph experiment results, the performance of KM-SVM algorithm was best.

4. Conclusion

The U-health system is for the user to be provided with health management service in real-time anytime, anywhere, specifically it is system for the feedback by the picking, storing, management, analysis on bio data in the ubiquitous environment. As the health became a centric value of society, the personalized medical services such as the medical service specialization and diversification etc. are demanded, and the development of u-Health system is accelerated.

Recently, the development of sensor network technology and establishment of ubiquitous environment became the practical use, the countless stream data measured from the user's several sensors are collected by U-health system in real time. The distribution of stream data may be changed, and a lot of data can be collected in short time, so the efficient energy storing, management are needed.

Therefore, this study placed a large number of sensor (blood pressure, pulse, body temperature, blood sugar) to the subjects, elderly over 65 years old(20 males, 32 meals), and the File System using the MapReduce framework was applied for the efficient input stream process, and the performance and cluster analysis was compared through the k-nn, K-means, SVM, KM-SVM algorithm. In the results, the cluster analysis on K-means algorithm showed better result than SVM algorithm, conversely, the performance of SVM algorithm was better. About KM-SVM, the performance and cluster analysis shows the results of combining the advantage of SVM algorithm and K-means algorithm. Through the KM-SVM algorithm, the cluster analysis effective data was used for U-health system. In the results of experiment by using 2003 data sets obtained from 52 test subjects, the k-NN algorithm showed 79.29% cluster analysis accuracy, K-means algorithm

showed 87.15 cluster analysis accuracy, and SVM algorithm showed 83.72%, KM-SVM showed 90.72%. As a result, the process speed and cluster analysis effective ratio of KM-SVM algorithm was better.

As the future direction of research, more efficient algorithm considering the cluster analysis and prediction will be developed, and study using MapReduce framework in many data in real time will be conducted for the process of data which was influenced by the flow of time. In addition, not only the sensor data in the limited laboratory, but also the study on the efficient process technique about the various vital information such as the user's location information, blood sugar, body fat etc. should be conducted, and the significant integrated expert monitoring system should be implemented to the judgment of the professional medical employee.

Acknowledgment

This study was supported by research funds from Chosun University, 2013.

References

[1] S.-D. Oh, “U-health system for efficient processing of multi-dimensional biological data stream”, M.S. Thesis, Chosun University, 2010.

[2] S.-H. Park, “Stream data splitting and allocation techniques for distributed parallel processing of real-time stream data”, M.S. Thesis. Pusan National University, 2013.

[3] S.-S. Yeo, H.-G. Yun, and S.-K. Kim, “For intellectual property protection of digital contents of anonymous fingerprinting research trends”, Korea Institute of Information Security & Cryptology, Vol. 11, pp. 90-99, 2001.