International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 1, February 2016, pp. 330~336
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i1.9334

Prediction Data Processing Scheme using an Artificial Neural Network and Data Clustering for Big Data

Se-Hoon Jung*, Jong-Chan Kim, Chun-Bo Sim*****
*Department of Multimedia Engineering, Sunchon National University,
(Gwangyang Bay SW Convergence Institute), Korea
** Department of Computer Engineering, Sunchon National University, Korea
*** Department of Multimedia Engineering, Sunchon National University, Korea

Article Info

Article history:

Received Jul 24, 2015
Revised Nov 12, 2015
Accepted Nov 28, 2015

Keyword:

Artificial neural network
Clustering
K-means
Principal component analysis
R programming

ABSTRACT

Various types of derivative information have been increasing exponentially, based on mobile devices and social networking sites (SNSs), and the information technologies utilizing them have also been developing rapidly. Technologies to classify and analyze such information are as important as data generation. This study concentrates on data clustering through principal component analysis and K-means algorithms to analyze and classify user data efficiently. We propose a technique of changing the cluster choice before cluster processing in the existing K-means practice into a variable cluster choice through principal component analysis, and expanding the scope of data clustering. The technique also applies an artificial neural network learning model for user recommendation and prediction from the clustered data. The proposed processing model for predicted data generated results that improved the existing artificial neural network–based data clustering and learning model by approximately 9.25%.

Corresponding Author:

Chun-Bo Sim,
Department of Multimedia Engineering,
National Sunchon University,
Maegok-Dong, Suncheon-si Jeollanam-do 540-742, Republic of Korea,
Email: cbsim@sunchon.ac.kr

1. INTRODUCTION

Today, real-time data and documents are on an exponential rise based on advanced mobile computing and social networking sites (SNSs). Created in real time, big data has atypical data structures, such as film and images added to the typical data structures created before now. Usually used by large corporations, big data analysis and prediction technologies are also utilized by government agencies, small and medium-sized companies, and today's common research institutions. There have been many studies on big data analysis and prediction technologies. The techniques of predicting typical or atypical big data created in real time are divided into supervised and unsupervised learning. While supervised learning is a type of machine learning to infer certain functions from the data, comprised of predicted answers, unsupervised learning finds relations among data by making use of unlabeled data. The number of studies is rising in the area of utilizing an artificial neural network and unsupervised learning–based clustering techniques to analyze big data being created now [1]. Studies on unsupervised learning–based clustering techniques propose big data analysis and processing methods [2-3]. Recommendation services to meet the requirements of users, and time technology for analysis of clustering processing is growing in importance, along with the big data analysis technologies. There are three major stages of data analysis clustering in the previous studies [4]. First, the pre-processing stage presents a structure in which raw data reflected for analysis are in a structure comprised of sentences containing words. It eliminates stop words and extracts

morphemes from sentences. The second stage distinguishes the number of clusters to determine the clustering of sentences pre-processed in the first stage, and repeats clustering by calculating the Euclidian distance of pre-processed data objects. The last stage is a structure of proposed users' predicted clustering through clustered data objects and provides fast operation speed for data analysis. Previous studies on data analysis and prediction had fixed and variable problems. First, they have to fix the determination coefficient of a cluster when the sample data for analysis increases, which is a disadvantage. In such a case, unnecessary data objects can be clustered unless the desired clustering happens. Secondly, they lacked the accuracy and reliability of prediction clustering by fixing certain labels, such as supervised learning, in advance, and clustering data to the fixed labels in low-level data analysis. The present paper proposes C- and R-based data result prediction performance analysis with a data processing model to analyze and predict connections and rules among data based on users' big data. The main goals of the proposed prediction data processing model are to overcome the problem of determining the number of clusters in the stage before data clustering, and to increase the accuracy and reliability of prediction data to make decisions for various prediction processes. It processes users' prediction data, including inter-data regularity and main topics pursued by users through principal component analysis (PCA) and K-means algorithms based on the sentences written on the users' SNSs. The distinguished regular data objects propose user prediction clustering through repetitive learning by applying an artificial neural network. Users' sentences on an SNS can express the words used in the sentences as a vector in the characteristic multiple-dimension vector space.

2. RELATED WORK

The genetic and neural network algorithms are good examples of algorithms trying to translate what man actually learns into the computer as it is. The neural network consists of nodes with mathematical computation capabilities interconnected to each another, operating by proper learning rules [5-6]. That is, each node performs a mathematical operation through coupling and transfer functions. The signals actually entered into the nodes can be expressed like Formula (1) based on the addition of weighted values and transfer functions.

$$S_j = \sum_{i=0}^{m} w_{ij} x_i \quad (1)$$

S_j, an actual input signal, obtains output values by going through a non-linear function called a transfer or activation function. w_{ji} represents the connection strength between input and hidden layers. x_i is the input value of an input layer. The sigmoid function is the most commonly used non-linear function, and multi-layer artificial neural networks consist of input, hidden, and output layers. Applied to the learning algorithm used to optimize connection strength was the error back propagation learning algorithm based on the gradient descent method, to which momentum constants and learning rates were applied. Included in the algorithm, an activation function used a tangent sigmoid function in the hidden layer and a linear function in the output layer. The gradient descent method repeatedly explores the optimization of parameters to improve the values of objective functions by calculating the adjustment volume in proportion to a primary derived function–based gradient of the objective function and improving the values of the objective function. Formula (2) shows an artificial neural network that used momentum constants and learning rates to provide more efficient training and to produce better results: γ is the learning rate; α the momentum constant; $w_{ji}^{(l)}(n)$ the connection strength connected to each layer; and $y_i^{(l-1)}(n)$ is the error rate.

$$w_{ji}^{(l)}(n+1) = w_{ji}^{(l)}(n) + \alpha\left[w_{ji}^{(i)}(n-1)\right] + \gamma\delta_j^{(l)} y_i^{(l-1)}(n) \quad (2)$$

3. PROPOSED DATA PROCESSING SCHEME

3.1. The Structure of the Data Processing Scheme

As seen in Figure 1, the proposed decision making clustering consists of extracting clusters through pre-processing, PCA, and a K-means algorithm, clustering analysis objects, and testing objects through an artificial neural network.

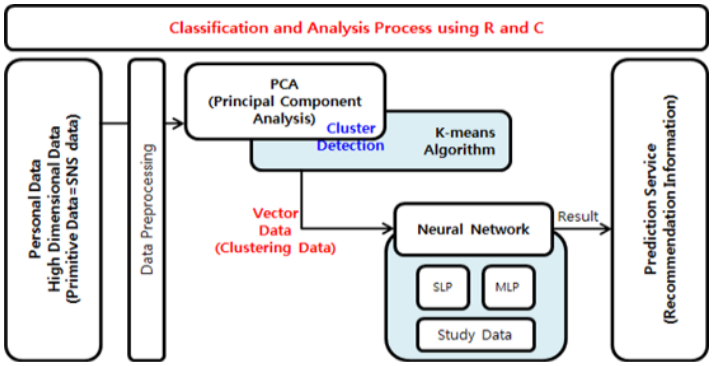

Figure 1. Overall structure of our scheme

The pre-processing stage produces input data by eliminating stop words and extracting morphemes from raw SNS data. Pre-processed data consist of a document–word matrix. The number of clusters is searched by conducting PCA for the characteristic vectors corresponding to the matrix, and users' past acts are clustered by applying the K-means algorithm. Pre-processed data are applied as learning data for data reliability and accuracy, undergo a clustering test process, and check the output results of the final clustered data.

3.2. Pre-Processing for the Data Object

This is a data pre-processing stage to apply raw SNS data to decision making clustering, extracting morphemes and eliminating stop words from SNS sentences that are raw data. The frequency of words in sentences is used as an important indicator to measure importance. If words have too high a frequency, however, they will decrease in importance and eventually hold no significance. The present paper applied algorithms to analyze morphemes and eliminate stop words by morpheme to extract morphemes from SNS sentences [7]. Figure 2 presents the conditions of stop words according to morphemes in the algorithm. *Minrangesupport* and *Mindelta* refer to the minimum scope support and its allowable range, respectively. If the number of morphemes to satisfy *Mindelta* is larger than the threshold, words will be extracted. If *Minrangesupport* is bigger than the mean of *rangesupport*, they will be processed as stop words.

$$Minrangesupport \leq wordsupport \quad (1)$$

$$k(|\Delta rangesup. * w_i(\alpha)| \leq Mindelta) \geq threhold \quad (2)$$

$$Minrangesupport > \frac{1}{k}\sum_{i=0}^{k}(rangesupport * w_i(\alpha)) \quad (3)$$

Figure 2. Condition for stop word detection

When, for instance, analyzing two sentences that a user has posted on an SNS ("I went to an amusement park for a date with my boyfriend today" and "I loved cotton candy the most at the amusement park today") with the levels classified by sentence days, morpheme analysis will produce results of {I, went, to, an amusement park, for, a date, with, my boyfriend, today, I, loved, cotton candy, the, most, at, the amusement park, today}. In morpheme analysis, the parts that are endings are all deleted to tell them apart from stems. When threshold is fixed at 1 with the data objects produced through morpheme analysis applied to a stop word processing algorithm, {today, boyfriend, data, today, amusement park, cotton candy, love} are extracted as candidates for *Minrangesupport*. Since the mean of *rangesupport* in all the sentences is 1, {today} is extracted as a stop word. There are a total of six SNS sentence words through morpheme analysis and stop word processing, namely {boyfriend (w1), date (W2), today (w3), amusement park (w4), cotton candy (w5) and love (w6)}. They are further divided into a day–word matrix by granting selective weights.

3.3. PCA and K-Means Algorithm for Clustering

Previous studies selected the number of clusters randomly or according to the number of cases wanted by the user in the early stage when making use of the K-means algorithm for data analysis[8-9]. If the number of vectors (the clusters for clustering) is chosen according to flexible and variable situations, it will be possible to predict data of broader scopes. Principal components of new data objects not correlated with

one another are extracted by changing linearly the multi-dimensional data of characteristic vectors based on the sentence–word matrix from the pre-processing stage. Extracted principal components are then used as vectors for clustering, which are processed according to the number of clusters. As the data analyzed through PCA show, SNS users increase in SNS utilization on Sunday, Monday, Thursday, Saturday, and Friday. They are applied to input data through the central point of each day vector. The initial values are determined around each cluster for the data objects according to the vectors selected based on the days analyzed as being principal components. Euclidian distance is obtained between the clusters analyzed as being principal components, and the data objects of SNS users for the data objects classified according to the day–word matrix. Once it is determined which data object has the highest similarity to which cluster, it will be moved to the cluster of concern. Finally, the central point of the moved cluster is re-calculated. Here is an example: for a data object called {Data}, distance is calculated based on the central point of each cluster and Euclidian distance, and the data object is moved from the cluster containing the minimum value to another cluster.

3.4. Design of Prediction Data ANN Model

A model is designed based on the data objects from cluster sentence data of SNS users. An artificial neural network model is built based on the trial-and-error method to predict users [5]. The artificial neural network constructs data objects classified in the pre-processing stage as input data and compares the results processed in the output layer with the performance of data objects produced via K-means. Formula (3) is a model equation to build an artificial neural network. Here, $T_{prediction}$ is the sum of recommended words associated with certain data objects; A_{ik} is an object that clusters with the K-means algorithm; and $n_{ik}$is a preliminary group of data objects classified through PCA, rather than predicted clusters.

$$T_{prediction}\sum_{k=0}^{n}[t_k] = ANN\begin{pmatrix} A_i,\ldots,A_k \\ n_i,\ldots,n_k \end{pmatrix} \quad (3)$$

```
Start ANN learning
 input data()=clustering object
 error data initial
begin
  for k do
   k < clustering data object count
   trial and error method
   compute direction = (middle weight, output weight, learning data set)
            input layer weight value addition
            threshold processing
        for i do
            hidden layer weight value addition
            return to sigmoid function
        end
    compute error calculation
        for b do
            connection strength learning
            threshold learning
        end
   compute input layer error value addition
     end
 end
 end
```

Figure 3. Artificial neural network learning algorithm for cluster analysis clustering

Figure 3 presents a learning algorithm for artificial neural networks to recommend associated words to the data objects produced through clustering. The learning algorithm generates learning data based on the word objects recommended through the day clusters produced through K-means. The trial-and-error method is used to calculate the sum of weight values and thresholds of word objects in the input layer. Errors are calculated through weight learning in the output layer. The learning and threshold learning of each connection layer are processed. Finally, error rates calculated repeatedly are added together to check the error rate of the final prediction object and judge the fitness of the recommended object.

4. EXPERIMENT AND PERFORMANCE EVALUATION

The proposed prediction data processing process was subjected to experiment and evaluated for performance in the following environments: the CPU was an Intel Core i7-4790 at 3.6GHZ with 16GB memory, and the O/S was Windows 7. The experiment tools used in the paper were Visual Studio and R Studio. Experiment data were constructed with 100 random SNS sentences posted by certain individuals over 14 days in order to analyze and predict their SNS data. The present paper conducted several experiments to assess the efficiency of the proposed prediction data processing from various perspectives. Nouns were extracted by eliminating stop words and endings in the pre-processing stage based on the raw SNS data. PCA

was carried out to obtain seeds, a cluster classification criterion for clustering, dynamically based on the extracted nouns. The SNS data objects were clustered with a cluster as a central point, and objects were selected for recommendation to users. Table 1 presents the day–word analysis matrix for PCA.

Table 1. PCA Result of Day–Word Matrix

Variable	PC1	PC2	PC3	PC4	PC5	PC6
sun.	-4.71	4.77	-4.41	-5.60	-0.89	-1.13
mon.	6.24	7.02	1.32	3.34	1.29	-3.10
tue.	-5.95	-3.72	-0.45	5.87	-1.81	-4.02
wed.	4.17	-2.62	-4.05	2.89	-1.02	5.03
thu	2.38	-5.74	0.64	-3.80	7.54	-2.13
fri.	2.81	-2.25	4.70	-3.71	-6.04	0.08
sat.	-4.95	2.57	5.24	1.00	2.93	5.27

The analysis results reveal that they were delivered in four vector numbers: *mon., tue., thu.,* and *sat.* Figure 4 shows the analysis results of relations between 119 input data (words) produced in the pre-processing stage and principal components. The results seen in Figure 4 cover the PCA results of PC1 and PC2 from a total of six rounds of PCA. Four principal components by day were identified, which include Monday, Tuesday, Thursday, and Saturday, when the users most often used the SNS. According to the evaluation results of SNS sentences for 14 days, the most influential words were {summer, shopping, telegram, and phone call}. Figure 5 presents an analysis graph by day for prediction data clustering. {Thursday, Saturday, Tuesday, and Monday} correspond to {black, green, red, and black}, respectively, falling into the centroid of each cluster on the clustering graph. The 119 words, which are input data objects, were classified into their clusters of concern by calculating the central location of each cluster and the Euclidian distance. The user recommendation results can ultimately be inferred for predictive service factors through pre-processing, PCA, and K-means algorithm analysis from the three following perspectives. First, it is possible to analyze the days when the SNS is used by the subjects; secondly, it is possible to check the areas of interest by day among the users; and finally, it can check the connections between the 119 words arranged through PCA and the optimized terms in the fields of the users' interests.

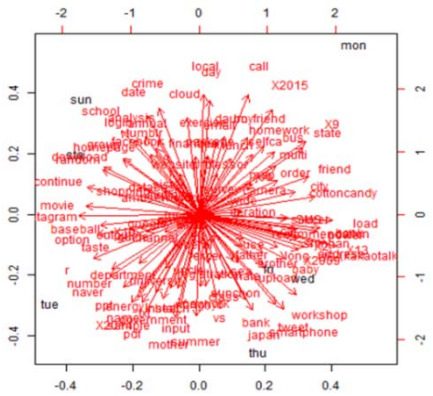

Figure 4. Input data analysis by PCA

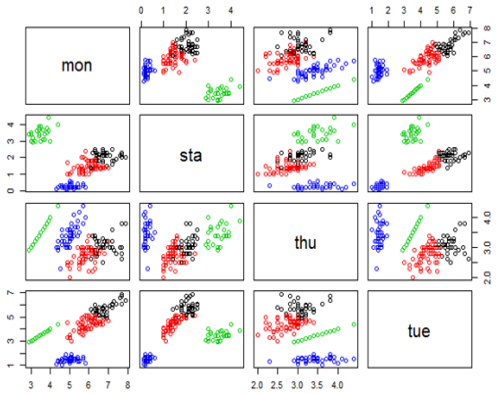

Figure 5. Input data clustering by K-means

The goal of an artificial neural network stage is to judge whether the classification prediction of a user recommendation service for certain data objects clustered through K-means analysis is accurate. For that goal, the present paper measured the predictions and error rates of artificial neural network–based data objects and checked the reliability of data prediction. The paper also checked the accuracy of {shopping} data objects in terms of analysis by day and predicted day. The learning number for the artificial neural network was fixed at 200 times in all cases. The learning results show that accuracy was the highest when the optimal number of hidden layers was two. Figure 6 presents the learning model creation results of an artificial neural network (ANN).

```
model.nnet <- nnet(mon ~ ., data = data.tr, size
= 2, decay = 5e-04, maxit = 200)
# weights:  35
initial  value 10.157702
iter  10 value 7.851735
iter  20 value 2.299154
.....
iter 170 value 0.286727
iter 180 value 0.261176
iter 190 value 0.242617
iter 200 value 0.214392
final  value 0.214392
stopped after 200 iterations
```

Figure 6. Results of creation by an ANN learning model

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}{n}}$$

$n : count\ number$
$x_{1i} : Actual\ Value$
$x_{2i} : Prediction\ Value$

Figure 7. Results of creation by an ANN learning model

Figure 7 is for accuracy and root mean squared error (RMSE), namely the numeric criteria to judge the efficiency and precision of a prediction data process. The measurement results are found in Table 2.

Table 2. PCA Result of Day-Word Matrix

{Shopping}	Node=2		Node=3		Node=4	
	Acc.	RMSE	Acc.	RMSE	Acc.	RMSE
sun.	86.7	0.8647	83.6	0.9451	84.7	0.7942
mon.	75.4	1.0148	64.6	1.3218	83.3	1.2115
tue.	88.8	1.1656	84.8	1.5444	81.2	0.8651
wed.	81.4	0.9947	79.7	0.8562	84.1	0.8451
thu.	85.5	0.7455	84.4	0.9186	77.7	0.7953
fri.	90.4	1.2432	79.9	0.7887	79.2	0.8844
sat.	93.8	0.9464	87.6	0.9499	85.4	0.9319

Table 3. Clustering Error Rate Measurement

	Data Clustering and Prediction (Only ANN)	Proposed Data Clustering and Prediction
Error Rate	14.81%	5.56%

Based on the learning measurements of the artificial neural network, the probability of each day for the {shopping} data objects was measured. The results indicate that Saturday held the highest measurements, regardless of the number of hidden nodes. The simple measured probability results agreed with data clustering through PCA and K-means. In addition, RMSE, which shows differences between predicted and actual values, had no big differences from simple probability results, recording 0.9464 when there were two nodes. Table 3 presents comparison results between the old artificial neural network–based data clustering and pattern prediction error rates, and those of the proposed data prediction model. The present paper had the same characteristics of unsupervised learning as previous studies, but the previous studies had to set classification criteria for clustering in advance. Because of the problem, the error rate of data object clustering or prediction was measured as high. The proposed paper generated effects of pre-processing the classification criteria through PCA and reduced the error rate by 9% or more in cases of data clustering and prediction under the same conditions.

5. CONCLUSION

In this paper, we proposed study of a processing model to cluster data based on user information created on an SNS and to predict about, and recommend to, users in the future. The proposed research model processes the clustering of user sentence data by making use of regularity among data pursued by those users, performing PCA on main topics, and applying a K-means algorithm based on the sentences users post on an SNS. The distinguished regular data objects provide user prediction data through repetitive learning by applying an artificial neural network. Data clustering proposed in the paper overcomes the problem of determining the number of clusters before clustering for diverse prediction decisions and improves the error rate by approximately 9.25% compared to previous studies, in terms of prediction data accuracy and reliability. We will supplement the parts where efficiency drops in the pre-processing stage because the present study covers SNSs with small amounts of sentence data, and we will continue to investigate the learning phenomenon of data objects toward the central points of clusters during K-means algorithm–based data clustering.

ACKNOWLEDGEMENTS

This work was supported by Research Foundation of Engineering College, Sunchon National University. The research was supported by 'Software Convergence Technology Development Program', through the Ministry of Science, ICT and Future Planning (S0170-15-1079, S0170-15-1054).

REFERENCES

[1] H.J. Kim, S.Z. Cho, and P.S. Kang, "KR-WordRank : An Unsupervised Korean Word Extraction Method Based on WordRank", *Journal of the korean institute of industrial engineers,* vol. 40, no. 1, pp. 18-33, 2014.
[2] C.W. Kim, S. Park, "Enhancing Text Document Clustering Using Non-negative Matrix Factorization and WordNet", *Journal of information and communication convergence engineerings*, vol. 11, no. 4, pp. 241-246, 2013.
[3] S. Park and S.R. Lee, "Enhancing document clustering using condensing cluster terms and fuzzy association", *Journal of IEICE Transactions on Information and Systems*, vol. 94D, no. 6, pp. 1227-1234, 2011.
[4] T. Zhang, "BIRCH: an efficient data clustering method for very large databases", *SIGMOD '96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data,* 1996, pp. 103-114.
[5] G. Barko, J. Hlavay, "Application of an artificial neural network (ANN) and piezoelectric chemical sensor array for identification of volatile organic compounds", *Talanta*, vol. 44, no. 12, pp. 2237-2245, 1997.
[6] H.K. Palo, M.N. Mohanty, "Classification of Emotional Speech of Children Using Probabilistic Neural Network", *International Journal of Electrical and Computer Engineering*, vol. 5, no. 2, pp.311~317, 2015.
[7] K.H. Joo, W.S. Lee, "Document Clustering based on Level-wise Stop-word Removing for an Efficient Document Searching", *Journal of the korean association of computer education*, vol. 11, no. 3, pp. 67-80, 2008.
[8] S.S. Kim, "Variable Selection and Outlier Detection for Automated K-means Clustering", *Journal of Communications for Statistical Applications and Methods*, vol. 22, no. 1, pp.55~67, 2015.
[9] D. Napoleon, S. Pavalakodi, "A new method for dimensionality reduction using K-means clustering algorithm for High Dimensional Data Set", *International Journal of Computer Applications*, vol. 13, no. 7, pp.41~46, 2011.

BIOGRAPHIES OF AUTHORS

Se-Hoon Jung received his BSc and MSc in Multimedia Engineering from Sunchon National University in 2010 and 2012, respectively. Currently, he is a senior researcher with the research & development team, Gwangyang Bay SW Convergence Institute, South Korea. His research interests include data analysis and data prediction.
E-mail : iam1710@hanmail.net

Jong-Chan Kim received a BSc, MSc, and PhD in computer engineering from Chonbuk National University, South Korea, in 2000, 2002, and 2007, respectively. He was a senior research professor in the Automation and System Research Institute at Seoul National University in 2013. His current research interests are image processing, computer graphics, data analysis.
E-mail : seaghost.sunchon.ac.kr

Chun-Bo Sim received a BSc, MSc, and PhD in computer engineering from Chonbuk National University, South Korea, in 1996, 1998, and 2003, respectively. Currently, he is an associate professor with the Department of Multimedia Engineering, Sunchon National University, South Korea. His research interests include multimedia databases, ubiquitous computing systems, and big data processing.
E-mail : cbsim@sunchon.ac.kr