International Journal of Engineering & Technology, 7 (2.28) (2018) 320-324

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Quasi-optimality under pseudo f statistic in clustering data

Teruhisa Hochin [1] *, Yoshihiro Hayashi [2], Hiroki Nomiya [1], Morshed U. Chowdhury [3]

[1] Faculty of Information and Human Sciences, Kyoto Institute of Technology, Japan
[2] Research And Development Department, Nitto Seiko Co., Ltd., Japan
[3] School of Information Technology, Deakin University, Australia
**Corresponding author E-mail: hochin@kit.ac.jp*

Abstract

Pseudo F statistic is often used in deciding the number of clusters. A set of clusters having the largest pseudo F value is selected as the optimum set of clusters. This paper proposes the quasi-optimum set of clusters, whose pseudo F value is larger than those of other sets of clusters, whose numbers are around the number of clusters in the quasi-optimum set. The before and behind (BB) difference of pseudo F values is proposed to find the number of clusters in the quasi-optimum set. The relative BB difference of pseudo F values, which is the ratio of the BB difference of pseudo F values to the pseudo F value itself, is also proposed to find it when the pseudo F value severely varies. This paper shows some examples to demonstrate that the BB differences of pseudo F values and the relative ones work well in finding qua-si-optimum sets of clusters.

Keywords: *Clustering; Difference; Pseudo F Statistic; Quasi-Optimum; Relative Difference.*

1. Introduction

Advances of computer and network technologies enable us to get various information over the Internet and receive various services. Wikipedia is an example of an encyclopedia on the Internet [1]. Anyone can read and edit its contents. Question and answer sites such as Yahoo! Answers [2] contain various answers, which constitute huge amount of information resources. Video sharing websites accept various videos, and permit anyone to view them. We can purchase various goods through online shopping sites. These sites have big lists of goods and huge amount of purchase records. These are so-called big data [3], [4]. Big data has three major characteristics: Volume, Velocity, and Variety. It is very hard to treat them. Cloud computing derivers various computing services over the Internet [4]. It provides us various services, and releases us from using fixed terminals. We will easily be able to treat huge amount of data by using cloud computing. Moreover, everything is beginning to connect each other through the Internet. This Internet of Things (IoT) technology will change a lot of things [5], [6]. We can use and control various equipment over the Internet. We can get various information from the equipment connected to the Internet. IoT is also changing manufacturing industries. In cloud manufacturing [7], manufacturing devices are connected through the Internet, and are controlled over the Internet. Here, large amount of data goes through the Internet [3], [4]. We must derive useful information from large amount of data. Machine learning [8] enables us to derive useful information.

Machine learning is divided into two categories: supervised and unsupervised [8]. Supervised learning takes pairs of input and output, and produces some inferring function to estimate an output from an input. Unsupervised learning takes inputs, and derive the underlying structure of inputs. Labeled data are not required for unsupervised learning. It is useful when the answers for training are not available, or not known. Clustering is a major method of unsupervised learning. This paper focuses on clustering.

Clustering is divided into two groups: hierarchical and nonhierarchical [8], [9]. In hierarchical clustering, data items close to each other are grouped into a cluster. Data items and clusters are recursively grouped up into a larger cluster. This process is graphically drawn as a dendrogram. The number of clusters is often decided based on the distances of clusters. On the other hand, the number of clusters is usually specified a priori in nonhierarchical clustering. K-means [8], [9] is one of nonhierarchical clustering methods. It is, however, often difficult to decide the number of clusters a priori. X-means solves this problem by automatically deciding it [10]. X-means uses Bayesian Information Criterion (BIC) for evaluating the validity of clusters. There are some cluster validity indices [11], [12]. Pseudo F statistics is one of these indices [13]. The set of clusters having the largest pseudo F value is the optimum set of clusters. In many situations, pseudo F statistics works well. However, there are some situations that the largest pseudo F value does not indicate the proper split of clusters [9].

This paper proposes the quasi-optimum set of clusters. At this set of clusters, the pseudo F value is not required to be the largest, however it is generally larger than those of other sets of clusters, whose numbers of clusters are around the number of the quasi-optimum clusters. The before and behind (BB) difference of pseudo F values is proposed to find the number of the quasi-optimum set of clusters. In case if the pseudo F value severely varies, then relative BB difference of pseudo F values, which is the ratio of the BB difference of pseudo F values to the pseudo F value itself, may be effective.

The structure of the remainder of the paper is as follows. Section 2 describes the pseudo F statistic. Section 3 proposes the quasi-optimum set of clusters, the before and behind (BB) difference of pseudo F values, and the relative one. Section 4 demonstrates the application of the BB difference. Section 5 discusses its characteristics and the careful point to use it. Finally, Section 6 concludes the paper.

Copyright © 2018 Teruhisa Hochin et. al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Pseudo F static

Calinski and Harabasz proposed a statistic describing the cluster validity [13]. This is called the pseudo F statistics. It is the ratio of between-cluster variance to within cluster variance. It is defined by (1).

$$\text{Pseudo F} = \frac{\sum_{k=1}^{K} n_k \|z_k - z\|^2/(K-1)}{\sum_{k=1}^{K}\sum_{i=1}^{n_k} \|x_i - z_k\|^2/(N-K)} \quad (1)$$

Where N is the number of points, K is the number of clusters, n_k is the number of points in the cluster k, z_k is the centroid of the cluster k, and z is the centroid of all the points.

Large pseudo F values indicate that clusters are dense, and are separated one another. It is said that the number of clusters having the largest pseudo F value is the optimum number of clusters. An example of the plot of pseudo F values according to the number of clusters is shown in Fig. 1. The pseudo F value becomes the largest when the number of clusters is four. The points in this data set should be divided into four clusters.

Although it is said that the largest pseudo F value shows us the optimum number of clusters, these clusters are not always effective to us [9]. Pseudo F values of various clusters are plotted in Fig. 2 according to [9]. This is the data set of protein consumption of Europe [14]. This data set includes the consumption of red meat, white meat, eggs, milk, fish, cereals, starchy foods, nuts, and fruits and vegetables in twenty-five countries. Clusters are obtained by using the fuzzy c-means clustering [15], [16], [17]. A point is decided to be contained in a cluster when the point has the highest degree of belonging to the cluster. The pseudo F value is the largest peak at two as shown in Fig. 2. As described in [9], two clusters are too few to get effective information from the clusters. In this plot, the pseudo F value of five clusters is the second largest peak when the pseudo F value of two clusters is considered to be the largest peak. Five clusters may give us more information than two clusters. Another criterion rather than the largest pseudo F value is required.

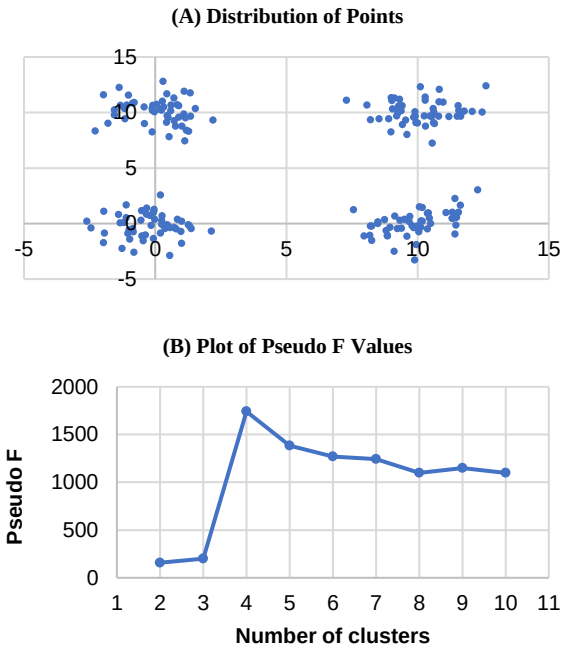


Fig. 1: Distribution of Points and the Plot of Pseudo F Values.

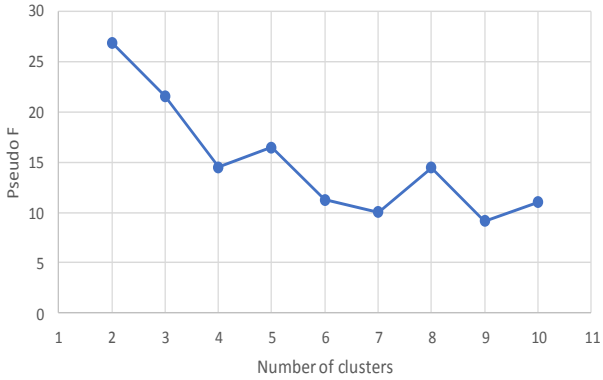


Fig. 2: Another Plot of Pseudo F Values [14].



Fig. 3: BB Difference of Pseudo F Values.

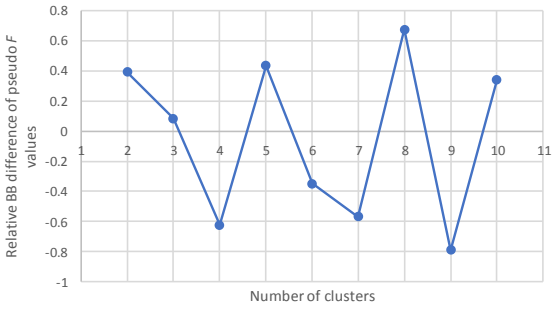


Fig. 4: Relative BB Difference of Pseudo F Values.

3. Quasi-optimum set of clusters

The difference of pseudo F values is proposed as a criterion in finding the number of clusters. This is based on the following observation: When pseudo F values are plotted according to the number of clusters, there are some peaks. For example, the plot shown in Fig. 2 has two additional peaks at five and eight. A peak at the number of clusters shows that the set of clusters is more optimum than those of clusters at before and after numbers of clusters. Such a set of clusters, however, may be less optimum than the set of clusters having the largest pseudo F value. Therefore, we may say such a set of clusters is quasi-optimum, whereas the set of clusters based on the optimum number is optimum. Peaks showing quasi-optimum sets can clearly be obtained by using the sum of the difference of the previous and the current pseudo F values, and that of the current and the next ones. The former (latter, respectively) difference is obtained by subtracting the previous (next) pseudo F value from the current one, i.e., $F_i - F_{i-1}$ (F_{i+1}). The sum of these differences is calculated by the formula $2F_i - (F_{i-1}+F_{i+1})$. From here on, this sum is called the before and behind (BB) difference of pseudo F values, and denoted as ΔF_i. The BB difference of pseudo F values is defined by (2).

$$\Delta F_i = \begin{cases} 2(F_2 - F_3) \\ 2F_i - (F_{i-1} + F_{i+1}) \\ 2(F_N - F_{N-1}) \end{cases} \quad (2)$$

The plot of the BB differences of pseudo F values shown in Fig. 2 is shown in Fig. 3. Peaks clearly appear. The BB difference of pseudo F values becomes a clue in finding the number of clusters in a quasi-optimum set. After we find the candidates of the numbers of clusters for the quasi-optimum sets, the best one can usually be decided according to the pseudo F value.

When pseudo F values severely vary, the value may become very large depending on the number of clusters, the BB difference of pseudo F values also becomes large according to the magnitude of the pseudo F value. In this case, it is difficult to choose which pseudo F value is the best for the quasi-optimum set. For showing the degree of influence (importance) of the difference, the ratio of the BB difference of pseudo F values to the pseudo F value ($\Delta F / F$) may be more effective than the pseudo F value itself. The ratio $\Delta F / F$ is called the relative BB difference of pseudo F values. The degree of the relative BB difference shows the degree of quasi-optimality of clusters. The ratio $\Delta F / F$ is obtained by dividing the BB difference of pseudo F values by the current one, i.e., $\Delta F_i / F_i$. For the BB differences shown in Fig. 3, the plot of the relative one is shown in Fig. 4. According to this plot, the division to eight clusters is more (quasi-)optimum, and may give us more information. Please note that this is an explanation of using the relative BB difference of pseudo F values. First, we should decide the best number of clusters according to the pseudo F value. In that case we could not decide it only by using the pseudo F value, the relative BB difference of pseudo F values should be used for deciding it.

4. Applications of quasi-optimum clusters

4.1. Points following gaussian distribution

Here, we examine two hundred two-dimensional points. Four sets of fifty points follow the Gaussian distribution, whose means are (0, 0), (2, 0), (0, 2), and (2, 2), respectively, and standard deviations are 1.0. The distribution of these points is shown in Fig. 5. C-means fuzzy clustering [15] is applied to them for obtaining clusters. Pseudo F values are calculated for each number of clusters. These are shown in Fig. 6. The pseudo F value becomes the largest at eight clusters. Those at four and nine are also large. The BB differences of pseudo F values are shown in Fig. 7. Although the BB difference becomes the largest at three clusters, the pseudo F value at four clusters becomes larger than that at three clusters. This phenomenon will be explained in Section V. The BB differences around eight are not so large. The relative BB difference of pseudo F values shown in Fig. 8 emphasizes this tendency. Although the pseudo F values are large around eight, the BB difference shows that the set of four clusters is quasi-optimum.

4.2. Realistic data

C-means fuzzy clustering [15] is applied to 11400 eight-dimensional point data. These are eight power spectrum coefficients of waveforms of signals which may include impulsive acoustic emission arising when objects are broken. The frequencies of the power spectrum coefficients are 9, 30, 60, 90, 120, 150, 180, and 210 kHz. These were obtained by the pre-experiments examining the effects of the power spectrum of the waveforms. Pseudo F values are calculated for each number of clusters. These are shown in Fig. 9. The pseudo F value becomes the largest at seven clusters. The BB differences of pseudo F values are shown in Fig. 10. Although the BB difference becomes the largest at seven clusters, the BB difference becomes large at three clusters. As the pseudo F values at three clusters and that at seven ones are quite different, therefore the relative BB differences of pseudo F values are calculated. These values are plotted and shown in Fig. 11. The relative BB difference of pseudo F values at three clusters is still not larger than that at seven clusters, but is relatively large. The set of three clusters is considered to be a quasi-optimum set of clusters. We may be able to select the set of three clusters as the final one.

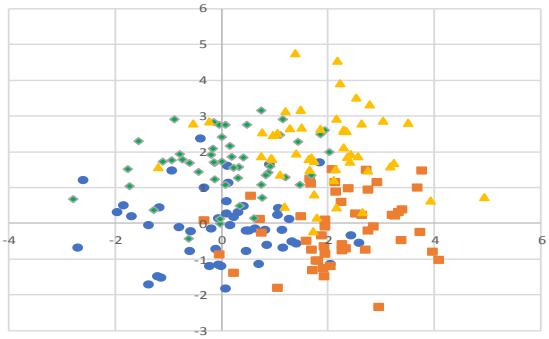


Fig. 5: Data Distribution.

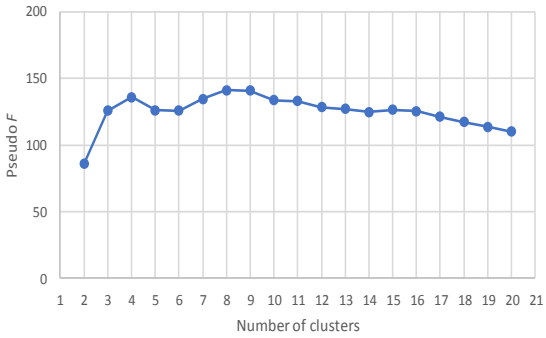


Fig. 6: Pseudo F Values.

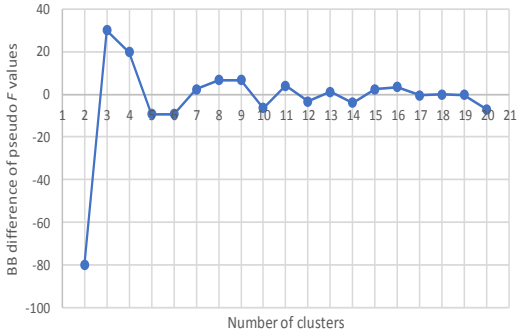


Fig. 7: BB Difference of Pseudo F Values.

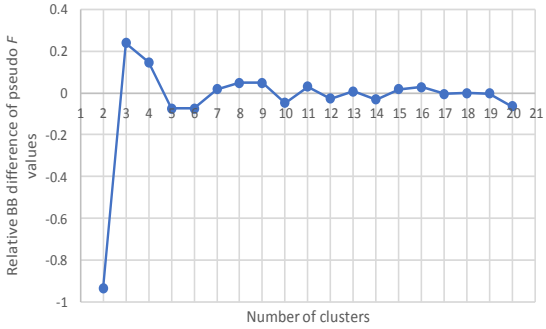


Fig. 8: Relative Bb Difference of Pseudo F Values.

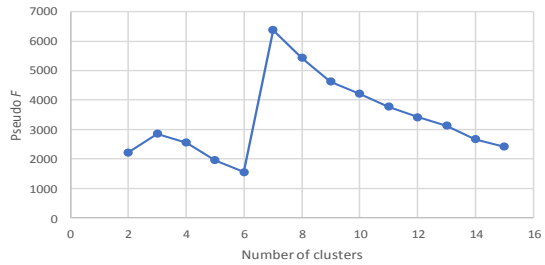


Fig. 9: Pseudo F Values.

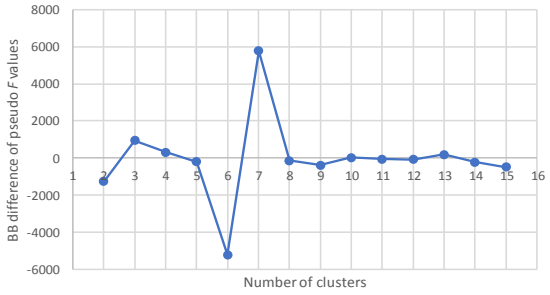


Fig. 10: BB Difference of Pseudo F Values.

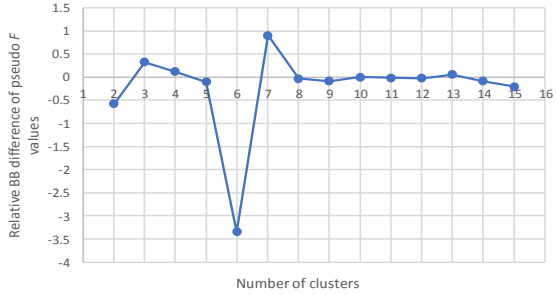


Fig. 11: Relative BB Difference of Pseudo F Values.

5. Discussion

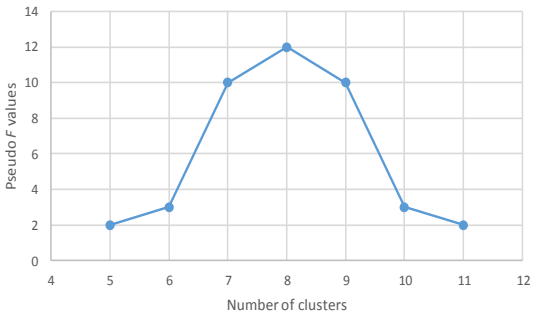
5.1. Quasi-optimum set of clusters

There are several clustering validity indices in the literatures [11], [12]. The set of clusters having the largest value of a clustering validity index is considered to be the optimum set of clusters. Although the clustering validity index indicates that some set of clusters is optimum, the set of clusters sometimes cannot sufficiently explain the characteristics of data. Clustering validity indices usually pay attention to only the largest value. The quasi-optimum sets of clusters are the sets of clusters having the values of a cluster validity index smaller than the largest one. The concept of the quasi-optimum set of clusters opens the new way of deciding the appropriate number of clusters. The appropriate set of clusters can be decided from the quasi-optimum sets of clusters as well as the optimum one.

5.2. BB difference

The BB difference of pseudo F values corresponds to the height of a peak around it. The BB difference is calculated by $(F_i - F_{i-1}) + (F_i - F_{i+1})$. This means the twice of the height of the peak. It is considered that the BB difference of pseudo F values is reasonable to detect a peak of pseudo F value.

(A) Pseudo F Values



(B) BB Difference Of Pseudo F Values

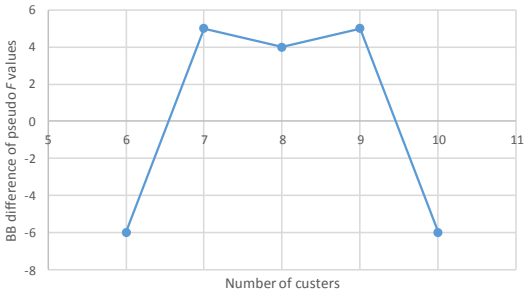
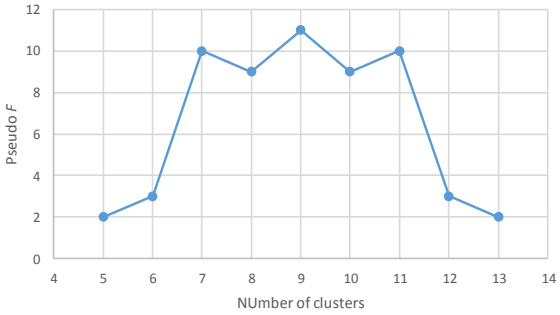


Fig. 12: Case That Pseudo F Values form A Hill.

(A) Pseudo F Values



(B) BB Difference of Pseudo F Values

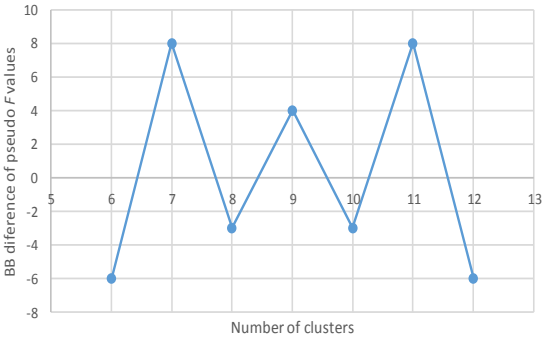


Fig. 13: Case That Pseudo F Values form a Crown

The BB difference of pseudo F values explains us the local importance of some set of clusters. Therefore, it could indicate quasi-optimum sets of clusters, which are kinds of local maximum of pseudo F values.

5.3. Usage of BB difference

In some condition the BB difference of pseudo F values does not work well as shown in Fig. 12. The pseudo F values shown in Fig. 12 (a) form a hill rather than a peak. The BB differences are large

at the hill sides, which are at seven and nine. Although the pseudo F value at eight is the largest, the BB difference shown in Fig. 12 (b) cannot directly indicate this number. We should consider that the BB differences catch the hill sides rather than the peak sides. We should search the points between two hill sides to find the point having the largest pseudo F value.
Another similar condition is shown in Fig. 13. Here, pseudo F values form the shape of a crown. This is a special case of a hill. Some pseudo F values between a crown sides are smaller than those of the tops of crown sides. In this case, the BB difference of pseudo F values at the point where the pseudo F value is the largest as well as those at the points of the tops of crown sides become large, whereas the value of the BB difference at the optimum number is smaller than those at the tops of crown sides. This condition may be easier than the previous one in finding the quasi-optimum point.

5.4. Relative BB difference

Relative BB difference shows the degree of influence of the difference. This is considered to be effective when the values of peaks are quite different. There are two peaks in the plot shown in Fig. 9. The values are about 2800 and 6400, respectively. Their BB differences are about 900 and 5800. Although these are quite different, the values of the peaks are also quite different. In this condition, using the relative BB difference may be effective because we should consider the magnitude of the values of the peaks.
On the other hand, the values of the peaks in the plot shown in Fig. 6 are about 135 and 140, respectively. As these are comparable, the BB differences work well. In this case, we do not have to use the relative BB differences.
In which condition the relative BB difference should be used is not clear. We must address this issue in our future work.

6. Concluding remarks

This paper proposed the quasi-optimum set of clusters, whose pseudo F value is not larger than the largest one. The pseudo F value of such a set of clusters is larger than those around the number of clusters in the set. The before and behind (BB) difference of pseudo F values was proposed to find the quasi-optimum set of clusters. The relative BB difference, which is the ratio of the BB difference of pseudo F values to the pseudo F value, was also proposed for the case that pseudo F values severely vary. It is shown that the BB difference works well in finding quasi-optimum sets of clusters. It is also shown that we should pay attention to the situation that pseudo F values form a hill rather than a peak or crown.
Direct usage of the BB difference of pseudo F values is hard for the situation that pseudo F values form a hill or the shape of a crown rather than a peak. Clarifying the precise procedure of finding the quasi-optimum number of clusters even in this situation is in future work. This may result in the automatic detection of the quasi-optimum number of clusters. Clarifying the condition that the BB differences of pseudo F values should be used is also in the future work.

References

[1] Wikimedia Foundation, "Wikipedia," https://en.wikipedia. org/.
[2] Yahoo Group, "Yahoo! Answers," https://answers.yahoo. com/.
[3] S. Sagiroglu and D. Sinanc, "Big Data: A Review," *International Conference on Collaboration Technologies and Systems*, (2013), pp. 42-47.
[4] D. Agrawal, S. Das, and A. E. Abbadi, "Big Data and Cloud Computing: Current State and Future Opportunities," *Proceedins of the 14th International Conference on Extending Database Technology (EDBT/ICDT '11)*, (2011), pp. 530-533.
[5] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, Vol. 29, No. 7, (2013), pp. 1645-1660.
[6] I. Lee and K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," *Business Horizons*, Vol. 58, No. 4, (2015), pp. 431-440.
[7] W. He and L. Xu, "A state-of-the-art survey of cloud manufacturing," *International Journal of Computer Integrated Manufacturing*, Vol. 28, No. 3, (2015), pp. 239-250.
[8] S. Marsland, *Machine Learning*, Chapman & Hall/CRC, (2015).
[9] N. Zumel and J. Mount, *Practical Data Science with R*, MANNING, (2014).
[10] D. Pelleg, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," *Proceedings of the 17th International Conference on Machine Learning (ICML '00)*, (2000), pp. 727-734.
[11] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 12, (2002), pp. 1650-1654.
[12] [12] L. Wilkinson, L. Engelman, J. Corter, and M. Coward, "Cluster Analysis," http://cda.psych.uiuc.edu/multivariate_fall_2012/systat_ cluster_ manual.pdf (Accessed on Dec. 22, 2017).
[13] T. Calinski, and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, (1074), pp. 1-27.
[14] The Data and Story Library, http://lib.stat.cmu.edu/DASL/ Datafiles/Protein.html.
[15] Dept. of Electronics, Information and Bioengineering, Polytechnic University of Milan, "Fuzzy C-Means Clustering," https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/ cmeans.html (Accessed on Dec. 22, 2017).
[16] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, Vol. 3, (1973), pp. 32-57.
[17] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algoritms," *Plenum Press*, (1981).