An Automatic Labeling of K-means Clusters based on Chi-Square Value

2017 J. Phys.: Conf. Ser. 801 012071

(http://iopscience.iop.org/1742-6596/801/1/012071)

An Automatic Labeling of K-means Clusters based on Chi-Square Value

[1*]**R Kusumaningrum,** [2*]**Farikhin**

[*]Master Program of Information System, School of Postgraduate Studies
[2]Department of Informatics, Faculty of Science and Mathematics
[3]Department of Mathematics, Faculty of Science and Mathematics,
Diponegoro University

E-mail: [1]retno@live.undip.ac.id, [2]farikhin.math.undip@gmail.com

Abstract. Automatic labeling methods in text clustering are widely implemented. However, there are limited studies in automatic cluster labeling for numeric data points. Therefore, the aim of this study is to develop a novel automatic cluster labeling of numeric data points that utilize analysis of Chi-Square test as its cluster label. We performed K-means clustering as a clustering method and disparity of Health Human Resources as a case study. The result shows that the accuracy of cluster labeling is about 89.14%.

1. Introduction

Clustering is an unsupervised method that is widely used in several domains to categorize data into some clusters based on its similarity. In image domain, clustering can be implemented as a method in image segmentation [1], images database categorization [2], image quality verification [3], etc. Moreover, clustering are also widely implemented in text domain, such as [4],[5],[6], etc.

In general, the studies of clustering method only implement related method for grouping the data into several clusters without annotations (automatic labeling) against the resulted clusters. Some of studies in text domain have been implemented cluster labeling, such as Wikipedia-based cluster labeling [7], hierarchical cluster labeling [8], [9], etc. However, there are limited studies for automatic cluster labeling in other domains. The existing cluster labeling for numeric data points is cluster labeling method for Support Vector Clustering (SVC) which is developed based on some invariant topological properties of a trained kernel radius function [10]. Its labels are support vectors, data points, and stable equilibrium points.

On the other hand, Chi-Square test is statistical methods that can be used detect different conditions between actual condition and ideal condition. Therefore, we proposed a novel method for cluster labeling of numeric data points that utilize analysis of Chi-Square test as its cluster label. In this study we performed K-means clustering as a clustering method and a disparity of Health Human Resources (HHR) as a case study.

2. Related Works

As clustering is an unsupervised method which categorizes data on some clusters based on its similarity, then there is no label for each resulted cluster. This leads to the emergence of challenges in the cluster labeled according to the data characteristics. The cluster label has been widely implemented in text data domain. However, it still very rarely implemented in numeric data. One of implemented methods in text data domain is chi-square method. It is used to test each word at each

node in a hierarchy starting at the root and recursively moving down the hierarchy [11]. Since chi-square method can be used to find the association between two variables, i.e. observed value and expected value, then it is possible to be implemented in numeric data for cluster labeling.

2.1. Clustering

There are several methods of clustering that are widely implemented such as K-means Clustering, K-Medoids, Hierarchical Clustering, DB Scan, etc. In this study, we apply K-means Clustering as its simplicity. We implemented K-means Clustering with the following specification, namely a standard *Euclidean distance* as similarity measure and Rule of Thumbs formula as a method for determining the number of cluster based on the following equation:

$$K \approx \sqrt{\frac{n}{2}} \tag{1}$$

Where K is defined as the number of cluster and n is the number of data.

The algorithm of K-means Clustering is as follow [12]:

1. Select k initial cluster centroid
2. Iteratively refining them as follows:
   a. Each instance d_i is assigned to its closest cluster centroid
   b. Each cluster centroid C_j is updated to be the mean of its constituent instances
3. Stop the iteration when there is no further change in assignment of instances to each cluster.

2.2. Cluster Homogeneity

Cluster homogeneity can be determined based on the silhouette coefficient value that can be obtained through the stages as follows [13]:

1. Calculate the average distance from a document supposes i with all other documents that are located in one cluster

$$a_i = \frac{1}{|A|-1}\sum_{j\in A, i\neq j} d(i,j) \tag{2}$$

   Where $|A|$ is defined as the number of data in Cluster A, $d(i,j)$ is defined the distance between i^{th} document and j^{th} document, and (i,j) is defined as the index of document
2. Calculate the average distance from i^{th} document with all the documents in the cluster and take the smallest value

$$d(i,C) = \frac{1}{|A|}\sum_{j\in C} d(i,j) \tag{3}$$

   Where $d(i,C)$ is defined the distance between i^{th} document over all object in other cluster, in which A ≠ C
3. The value of the Silhouette coefficient is

$$s(i) = \frac{b(i)-a(i)}{\max\big(a(i),b(i)\big)} \tag{4}$$

2.3. The Standard Analysis of Requirement Status of HHR

According to the Regulation Number 32 of 1996, health professionals is everyone who devoted themselves in the field of health and has the knowledge and/or skills through education in the field of health, which requires the authority to do the efforts of health in specific types. Based on the Regulation of the Minister of Health Republic of Indonesia, Number 33 of 2015 on Preparations Guidelines of Health Human Resource Requirement Planning, which is included into HHR, i.e. (i) health care practitioner includes general practitioner, dentist, and specialists; (ii) nursing includes nurses, dentist nurses, and midwives; (iii) pharmacy includes pharmacies and assistant pharmacists; (iv) public health officer; (v) sanitarian, (vi) nutrition experts, and (vii) therapists both physical and medical.

One of methods that can be applied to plan the HHR requirement is "population ratio" method, namely the ratio of health professionals to the total population for a region. The ratio of the standard population is determined based on population per regency, population growth per regency as well as a number of health professionals.

On the other hand, a disparity analysis can be implemented to determine the appropriateness between the availability of HHR with "standards of the population ratio". A Chi-Square analysis is a method that can be used to perform the disparity analysis with the following formula:

$$X^2 = \sum \frac{(F_o - F_h)^2}{F_h} \quad (5)$$

Where X^2 is *Chi-square* value, F_o is observed value, and F_h as predicted value

The interpretation value of the Chi-Square is based on the comparison between the values of arithmetic Chi-Square with the value of Chi-Square table. The greater value of arithmetic Chi-Square means there is a significant difference between the existence of health professionals with the normative needs of health care professionals, and it's applied on the contrary. Subsequently, this framework applied as a method for cluster labels.

3. Research Procedures

Illustration of the research procedure can be seen in Figure 4, whereas the details explanation about each step can be seen on the following subsection.

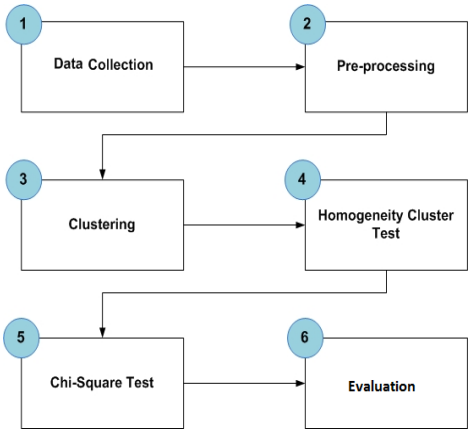

Figure 1. Research procedures

3.1. Data Collection

As explained before, the collected data including:

- Population data obtained from the Central Bureau of Statistics, Central Java Province
- HHR data (doctors, nursing, midwives) are acquired through an online media, i.e. dashboard of Health Human Resources Information System version 2016. This system is published by the PPSDM, Ministry of Health Republic of Indonesia.

In addition, this research also uses the HRR Requirement ratio per 100,000 total populations as described in Table 1.

Table 1. HHR Requirement over Population Data

No	Type of Health Professional	HHR Requirement per 100.000 Total Populations
1	General Practitioner	40
2	Specialist	10
3	Dentist	12
4	Nurse	158
5	Midwife	100

3.2. Pre-processing

The pre-processing step is used to normalize the obtained data into range [0,1] by using min-max normalization. This step is important since the range of the obtained data is difference for each type and it would give an impact on the performance of clustering process, where the process of the clustering data was conducted based on the similarity level between the data.

3.3. Clustering

The clustering process will be applied with K-means clustering method, whereas the Euclidean distance applied as the measurement techniques of similarity level. The detail of clustering methods has been described in Subsection 2.1.

3.4. Cluster Homogeneity Test

The aims on performing the cluster homogeneity cluster are to define the quality and the strength of the cluster, how good an object placed in a cluster. The method for performing the cluster homogeneity test on this research is Silhouette coefficient. The detail of the homogeneity cluster test can be seen on the Subsection 2.2.

3.5. Chi-Square Test

As described previously Chi-square test is performed to know the requirements disparity of HHR in accordance with the Government Regulation No. 32 of 1996 compared with the real conditions in the field about the availability of HHR.

3.6. Evaluation

Evaluation was performed to evaluate the accuracy level of the Chi-Square based labeling.

4. Result and Analysis

The results from the clustering process using K-means Clustering can be seen in Figure 2 with the value of the centroid for each cluster can be seen in Table 2 and the number of the cluster based on the Rule of Thumbs formula is defined as 4 clusters.

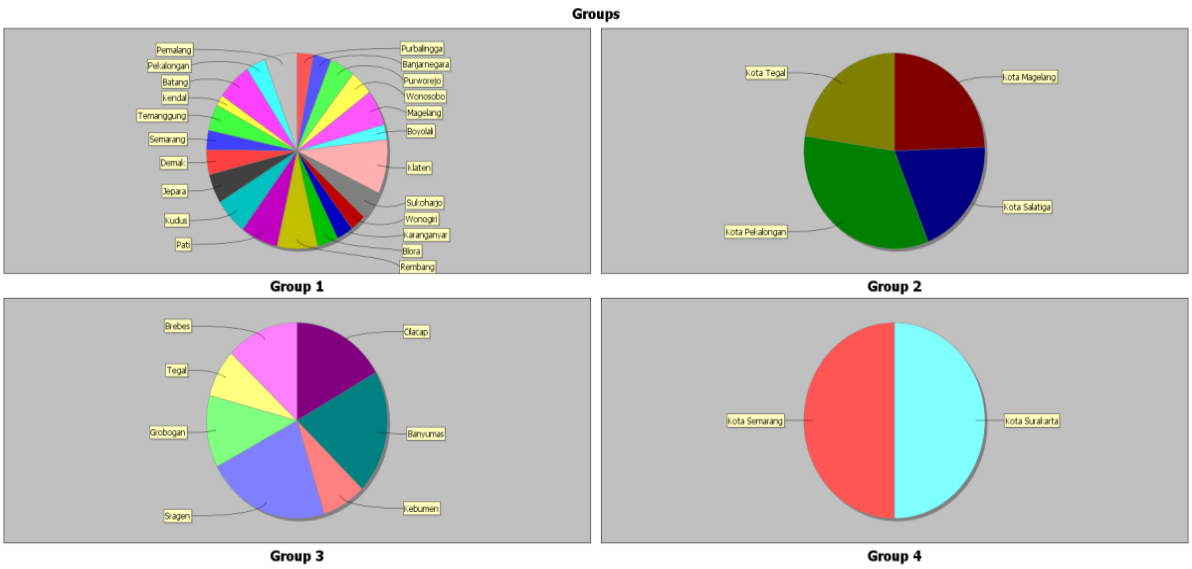

Figure 2. Clustering Result

Table 2. Last Centroid from *Clustering* Process

Cluster ID	The Number of Population	HHR Availability				
		General Physician	Specialist	Dentist	Nurse	Midwife
1	9363060	1160	990	310	6100	306
2	2101091	841	1231	231	5891	96
3	14143392	1482	1692	442	9512	662
4	10915493	3733	9283	1343	31043	404

As mention in the previous section, the cluster homogeneity test was performed based on Silhouette coefficient of about 0.541 which means the results of clustering have a good structure.

Subsequently, the calculation of Chi-Square Test was performed for automatic cluster labeling in general as well as the identification of HHR disparity status (the difference between the value of the availability and value of requirement) for regency in Central Java province in specific.

Table 3. *Chi-Square* Test

	Cluster ID	General Practitioner	Specialist	Dentist	Nurse	Midwife
F_o – Availability	1	1160	990	310	6100	306
	2	841	1231	231	5891	96
	3	1482	1692	442	9512	662
	4	3733	9283	1343	31043	404
F_o – Requirement	1	3746	937	1124	14794	9364
	2	841	211	253	3320	2102
	3	5658	1415	1698	22347	14144
	4	4367	1092	1310	17247	10916
Expected Value (F_h)	1	2453	963.5	717	10447	4835
	2	841	721	242	4605.5	1099
	3	3570	1553.5	1070	15929.5	7403
	4	4050	5187.5	1326.5	24145	5660
Chi-Square Value (Availability)	1	681.5528	0.728853	231.0307	1808.788	4242.366
	2	0	360.749	0.5	358.8123	915.3858
	3	1221.217	12.34776	368.5832	2585.411	6138.198
	4	24.8121	3233.373	0.205239	1970.694	4880.837
Chi-Square (Requirement)	1	681.5528	0.728853	231.0307	1808.788	4242.366
	2	0	360.749	0.5	358.8123	915.3858
	3	1221.217	12.34776	368.5832	2585.411	6138.198
	4	24.8121	3233.373	0.205239	1970.694	4880.837
Chi-Square	1	1363.11	1.46	462.06	3617.58	8484.73
	2	0	721.5	1	717.62	1830.77
	3	2442.43	24.7	737.17	5170.82	12276.4
	4	49.62	6466.75	0.41	3941.39	9761.67

Since we used 2 categories (availability and requirement) then the degree of freedom is *(2-1) = 1*. Based on its value and fault tolerance 0.5, thus the value of Chi-Square table is 3.841. Therefore, we perform the automatic labeling process based on the rule in eq. (6) and the result label in table 4.

$$disparity\_label = \begin{cases} Fair, if\ Arithmetic\ Chi - Square < 3.841 \\ High, otherwise \end{cases} \quad (6)$$

Table 4. General Label for Each *Cluster*

Cluster ID	General Practitioner	Specialist	Dentist	Nurse	Midwife
1	HIGH	FAIR	HIGH	HIGH	HIGH
2	FAIR	HIGH	FAIR	HIGH	HIGH
3	HIGH	HIGH	HIGH	HIGH	HIGH
4	HIGH	HIGH	FAIR	HIGH	HIGH

Based on Figure 2 and Table 4, there are four cities in Central Java Province that have good disparity label (fair status for both general practitioner and dentist availability), namely the Magelang City, Salatiga City, Pekalongan City, and Tegal City. Those cities are included in cluster 2. The final evaluation process was performed in order to measure the accuracy of the labeling process based on the ratio between the number of true predicted label and the total number of labels. The result shows that the labeling accuracy is about 89.14%.

5. Conclusions and Future Works

The automatic labeling of K-means cluster based on Chi-Square Tests was successfully applied with the accuracy is about 89.14%. Based on the resulted label, there are four cities in Central Java Province that have good disparity label (fair status for both general practitioner and dentist availability), namely the Magelang City, Salatiga City, Pekalongan City, and Tegal City.

The same strategy can be applied to all the regencies and cities in Indonesia and to assess the availability of other HHR's such as pharmacists, public health officer, sanitarian, etc. While to improve the accuracy of the value can be done by optimizing the K-means Clustering algorithm thus they have better Silhouette coefficient.

References

[1] Dhanachandra N, Manglem K and Chanu Y J 2015 Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm *Proc. of Eleventh Intl. Multi-Conf. on Information Processing-2015* pp 764–771
[2] Le Saux B and Boujemaa N 2002 Unsupervised Categorization for Image Database Overview *Proc. of Intl. Conf. on Visual Information System (VISUAL'2002)*
[3] Niemeijer M, Abràmoff M D and van Ginneken B 2006 Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening *Med. Image Anal.* **10** no. 6 pp. 888–898
[4] Yin J and Wang J 2016 A Text Clustering Algorithm Using an Online Clustering Scheme for Initialization *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016 pp. 1995–2004
[5] Xu T and Oard D W 2011 Wikipedia-based Topic Clustering for Microblogs *Proc. of the American Society for Information Science and Technology* pp. 1–10.
[6] Aggarwal C C, Zhao Y and Yu P S 2012 On Text Clustering with Side Information *Proc. of the 2012 IEEE 28th Intl. Conf. on Data Engineering* pp. 894–904
[7] Carmel D, Roitman H and Zwerdling N 2009 Enhancing Cluster Labeling Using Wikipedia in *Proc. of the 32nd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* pp. 139–146.
[8] Treeratpituk P and Callan J. 2006 Automatically Labeling Hierarchical Clusters *Proc. of the 2006 Intl. Conf. on Digital Government Research* pp 167–176
[9] Moura M F and Rezende S O 2007 Choosing a Hierarchical Cluster Labelling Method for a Specific Domain Document Collection *New Trends in Artificial Intelligence* pp. 812–823
[10] Lee J and Lee D 2005 An Improved Cluster Labeling Method for Support Vector Clustering *IEEE Trans. Pattern Anal. Mach. Intell.* **27** 3 pp. 461–464
[11] Popescul A and Ungar L H 2000 Automatic Labeling of Document Clusters [Online] Available: http://www.cis.upenn.edu/~popescul/Publications/labeling_KDD00.pdf
[12] Wagstaff K, Cardie C, Rogers S and Schroedl S 2001 Constrained K-means Clustering with Background Knowledge *Proc. of the Eighteenth Intl. Conf. on Machine Learning* pp. 577–584
[13] Rousseeuw P J 1987 Silhouettes : a graphical aid to the interpretation and validation of cluster analysis *J. Comput. Appl. Math.* **20** pp. 53–65