ISSN 2320-9984 04
9 772320 998015
Volume 1, Issue 4, December 2013 ISSN: 2320-9984 (Online)
International Journal of Modern Engineering & Management Research
Website: www.ijmemr.org

Study of Multiview Data Using Automated Two-Level Variable Weighting Clustering Algorithm

Piyush Tiwari
M. Tech. Research Scholar
Department of Computer Science & Engineering
Patel College of Science and Technology Bhopal
Bhopal (M.P.) [INDIA]
Email :piyush83mca@yahoo.co.in

Prof. Hitesh Gupta
Assistant Professor
Department of Computer Science & Engineering
Patel College of Science and Technology Bhopal
Bhopal (M.P.) [INDIA]
Email : hitesh034@gmail.com

Abstract—*Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both. If meaningful groups are the goal, then the clusters should capture the natural structure of the data. In some cases, however, cluster analysis is only a useful starting point for other purposes, such as data summarization. Whether for understanding or utility, cluster analysis has long played an important role in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. Both view weights and variable weights are used in the distance function to determine the clusters of objects. In the new algorithm, two additional steps are added to the iterative-means clustering process to automatically compute the view weights and the variable weights. We used two real-life data sets to investigate the properties of two types of weights in TW-k-means and investigated the difference between the weights of TW-k-means and the weights of the individual variable weighting method.*

Keywords:—*Data mining, clustering, MultiView learning, k-means, variable weighting*

1. INTRODUCTION:

Multiview data are instances that have multiple views (representations/variable groups) from different feature spaces. It is the result of integration of multiple types of measurements on observations from different perspectives and different types of measurements can be considered as different views. For example, the variables of the nucleated blood cell data [1] Clustering for Understanding Classes, or conceptually meaningful groups of objects that share common characteristics, play an important role in how people analyze and describe the world. Indeed, human beings are skilled at dividing objects into groups (clustering) and assigning particular objects to these groups. For example, even relatively young children can quickly label the objects in a photograph as buildings, vehicles, people, animals, plants, etc. In the context of understanding data, clusters are potential classes and cluster analysis is the study of techniques for automatically finding classes. The following are some examples:

2. INFORMATION RETRIEVAL

The World Wide Web consists of billions of Web pages, and the results of a query to a search engine can return thousands of pages. Clustering can be used to group these search results into a small number of clusters, each of which captures a particular aspect of the query. For instance, a query of "movie" might return Web pages grouped into categories such as reviews, trailers, stars, and

theaters. Each category (cluster) can be broken into subcategories (sub-clusters), producing a hierarchical structure that further assists a user's exploration of the query results. Clustering for Utility Cluster analysis provides an abstraction from individual data objects to the clusters in which those data objects reside. Additionally, some clustering techniques characterize each cluster in terms of a cluster prototype.

This is guided by a general 'philosophy of clustering', which involves considerations of how to define the clustering problem of interest, how to understand and 'tune' the various available clustering approaches and how to choose between them. All this should be driven by the way that the subject matter researchers connect the aim of clustering and their interpretation of concepts like 'similarity' and 'belonging together in the same class' to the choice and formal handling of the indicators involved. The main contribution of this paper is to show in detail how this can be done and what it entails, exemplary for the socio-economic stratification application but including aspects, particularly in Section 6, that are relevant for other clustering applications as well and hardly mentioned in the literature. The present approach brings statistical and sociological (or general subject matter) knowledge closer together by focusing on the interface, the 'translation task' between statistical methodology and sociological background. Two quite different approaches are compared, namely a model-based clustering approach [2] in which different clusters are modelled by underlying latent classes or mixture components, and a dissimilarity-based partitioning approach that is not based on probability models [3] with some methods to estimate the number of clusters. Such data typically arise in social stratification and generally in social science. Social stratification is about partitioning a population into several different social classes. Although the concept of social class is central to social science research, there is no agreed definition of a social class. It is of interest here whether social stratification based on formal clustering can contribute to the controversy about social class. In this paper we analyse data from the US Survey of Consumer Finances (SCF), for which the more appropriate term is 'socio-economic stratification'. Apart from computing and interpreting a clustering, we address whether the clustering captures significant structure beyond decomposing dependence structures between the indicators, which indicators are most influential for stratification and whether strata derived from the data are related to occupation categories.

3. BACKGROUND

The concept of social class is central to social science research, either as a subject in itself or as an explanatory basis for social, behavioural and health outcomes. The study of social class has a long history, from the social investigation by the classical social thinker Marx to today's on-going academic interest in issues of social class and stratification for both research and teaching [4]. Researchers in various social sciences use social class and social stratification as explanatory variables to study a wide range of outcomes from health and mortality [5] to cultural consumption [6]

When social scientists employ social class or stratification as an explanatory variable, they follow either or both of two common practices, namely using one or more indicators of social stratification such as education and income and using some version of occupation class, which is often aggregated or grouped into a small number of occupation categories. For example, [7] compared health outcomes and mortality between white-collar and blue-collar workers[8] analysed the effects of social stratification on cultural consumption with a variety of variables representing stratification, including education, income, How to Find an Appropriate Clustering 311 occupation classification and social status (operationalized by them). The reason why researchers routinely use some indicators of social class is that there is no agreed definition of social class, let alone a specific agreed operationalization of it. Various concepts of social class are present in the sociological literature, including a 'classless' society [9], society with a gradational structure

[10] and a society in which discrete classes (interpreted in various, but usually not data-based, ways) are an unmistakable social reality [11] The question to be addressed by cluster analysis is to underst and 'social stratification' through data, i.e. to explore what kind of unsupervised classification(s) based on the social (or socio-economic) indicators they yield. We acknowledge that data alone cannot decide the issue objectively. Some questions to address are whether the data are meaningfully clustered at all, which indicators contribute most to a clustering and to what extent clusters are aligned with features that have been connected to social stratification in the literature, here particularly occupation.

Clusters may also serve as efficient reduction of the information in the data and as a tool to decompose and interpret inequality and changes over time. A latent class model was proposed[12]. A similar finite mixture model was proposed and applied to income inequality data [13].

Characteristics of Two Real-Life Data Sets

The data have been collected for the urban water, water treatment plant[14]. This data set contains 527 instances and 38 features. The 38 features can be naturally divided into four views. Input view contains the first 22 features describing different input conditions. The exit view contains the 23th-29th features de -scribing output demands. Actual working input views shows the 30th-34th features describing performance input demands. Global performance input view. shows the 35th-38th features describing global performance input demands. Here, we use G1; G2; G3, and G4 to represent the four views[15]. Further, it includes data for the B-type cyclin Clb2p and G1 cyclin Cln3p induction experiments. G1: contains four features from the B-type cyclin Clb2p and G1 cyclin Cln3p induction experiments;. G2: contains 18 features from the factor arrest experiment; G3: contains 24 features from the elutriation experiment; G4: contains 17 features from the arrest of a cdc15 temperature-sensitive mutant experiment;. G5: contains 14 features from the arrest of a cdc28 temperature-sensitive mutant experiment. In the following, we use the two real-life data sets to investigate the properties of two types of weights in TW-k-means.

Controlling Weight Distributions

From figure 2a, we can see that when ŋ was small, the variances of *V* were unstable with the increase of k. When ŋ was large, the variances of *V* became almost constant. From Figure.2b, we can see ŋ has similar behavior. While some different types of behaviors can be observed in the following figures that is Figure 2 c, d, e and f as shown in pictorial representation.

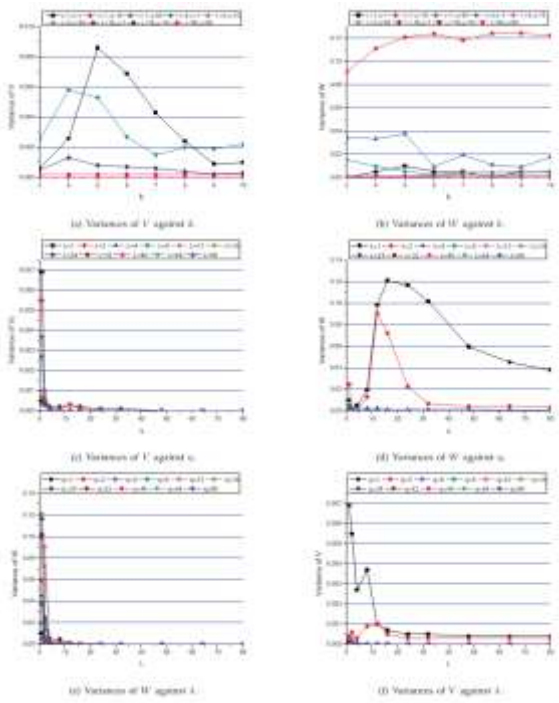
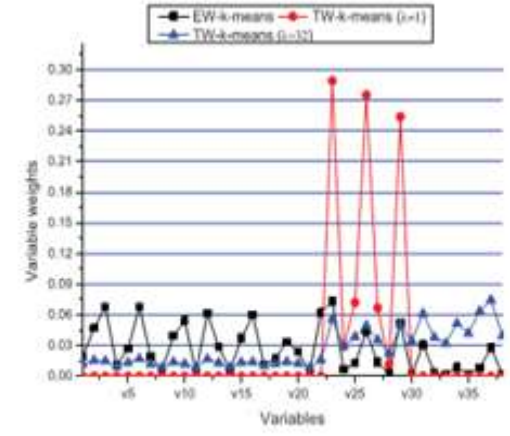


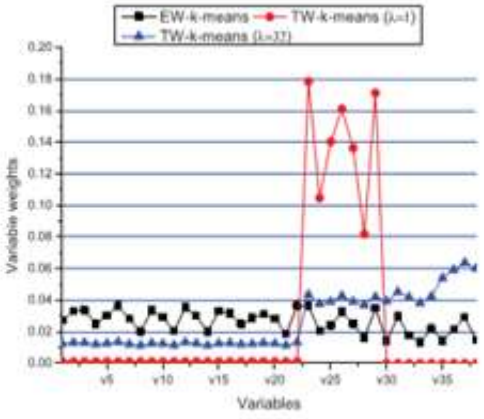
Figure 2: The variances of two types of weights *V* and *W* against three parameters *k*, 0 and 8 in TW-*k*-means on the Water Treatment Plant data set.

Comparison of the Weights in TW-k-Means and EW-k-Means

The data shown in figure 3 clearly show the variation of variables in both the axis that also represents the actual meaning of the term TW-k-Means and EW-k-Means [16,17].



(a) $\eta = 8$.

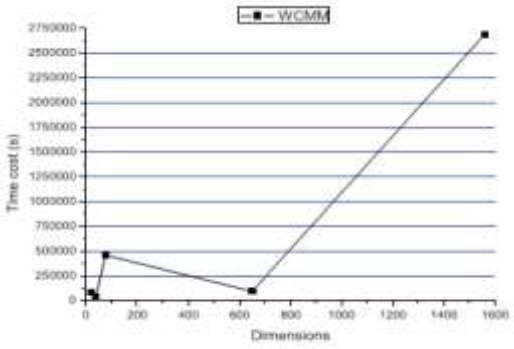


(b) $\eta = 32$.

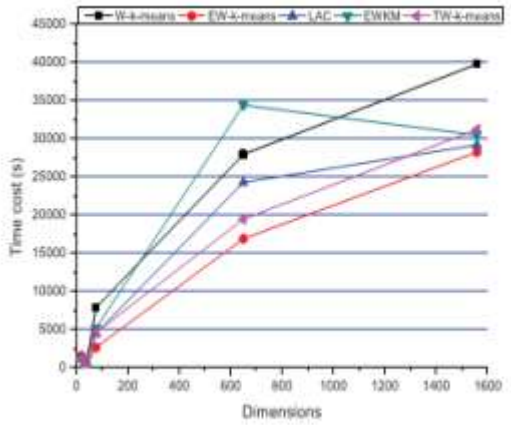
Figure 3: Comparison of the total variable weights in TW-*k*-means and EW-8-means 8and EW-8 means on the Water Treatment Plant data set.

Scalability Comparison

We used all five real-life data sets to compare the scalability of TW-k-means with the other five clustering algorithms.



(a) Average time cost of WCMM.



(b) Average time costs of the other five clustering algorithms.

Figure 4: (a) Average time cost of WCMM and (b) Average time cost of the other five clustering algorithms.

Figure 4 draws the average time costs of six clustering algorithms. We can see that the execution time of TW-k-means was only more than EW-k-means, and significantly less than the other four clustering algorithms. This result indicates that TW-k-means scales well to high-dimensional data [18-20].

4. CONCLUSION

First, the original data set is represented using a smaller set of prototype vectors, which allows efficient use of clustering algorithms to divide the prototypes into groups. There duction of the computational cost is especially important for hierarchical algorithms allowing clusters of arbitrary size and shape The experiments also revealed the convergence property of the view weights in TW-*k*-means. We compared TW-*k*-means with five clustering algorithms on three real-life data sets and the results have shown that the TW-*k*-means algorithm significantly outperformed the other five clustering algorithms in four evaluation indices. As such, it is a new variable weighting method for clustering of multiview data.

REFERENCES:

[1] J. Mui and K. Fu, "Automated Classification of Nucleated Blood Cells Using a Binary Tree Classifier,

”IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 2, no. 5, pp. 429-443, May 1980.

[2] Vermunt, J. K. and Magidson, J. (2005) Technical Guide for Latent GOLD 4.0: Basic and Advanced. Belmont: Statistical Innovations.

[3] Kaufman, L. and Rousseuw, P. J. (1990)Finding Groups in Data. New York: Wiley.

[4] Gower, J. (1968) Adding a point to vector diagrams in multivariate analysis.Biometrika,55, 582–585.

[5] Graf, S. and Luschgy, H. (2000) Foundations of Quantization for Probability Distributions. Berlin: Springer.

[6] Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F. and Pirahesh, H.

[7] (1997)Datacube: a relational aggregation operator generalizing group-by, cross-tab and sub-totals. DataMinng Know l. Discov.,1, 29–53.

[8] Irigoien, I. and Arenas, C. (2008) INCA: new statistic for estimating the number of clusters and identifying atypical units. Statist. Med.,27, 2948–2973.

[9] Irigoien, I., Fernández, E., Vives, S. and Arenas, C. (2008) Clum: a cluster program for analyzing microarray data. Russ. J. Genet.,44, 993–996

[10] McLachlan, G. J. (2011) Commentary on ‘Evaluating mixture modeling for clustering: recommendations and cautions’ by D. Steinley and M. J. Brusco. Psychol. Meth.,16, 80–81.

[11] Pollock, G. (2007) Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. J. R. Statist. Soc.A,170, 167 –173.

[12] Stehlík, M. (2003) Distributions of exact tests in the exponential family.Metrika,57, 145–164

[13] Roever, C. and Szepannek, G. (2005) Application of a genetic algorithm to variable selection in fuzzy clustering. In Classification—the Ubiquitous Challenge (eds C. Weihs and W. Gaul), pp. 675–681. New York: Springer.

[14] A. Frank and A. Asuncion, “UCI Machine Learning Repository,” http://archive.ics.uci.edu/ml, 2010.

[15] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher, “Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization,” Molecular Biology of the Cell, vol. 9, no. 12, pp. 3273-3297, 1998.

[16] N. Kushmerick, “Learning to Remove Internet Advertisements,” Proc. Third Ann. Conf. Autonomous Agents, pp. 175-181, 1999.

[17] Wikipedia, “Plagiarism—Wikipedia, the Free Encyclopedia,” http://en.wikipedia.org/wiki/Information_retrieval, 2011

[18] H. Cheng, K.A. Hua, and K. Vu, “Constrained Locally Weighted Clustering, ”Proc. VLDB Endowment, vol. 1, pp. 90-101, Aug. 2008.

[19] W. DeSarbo, J. Carroll, L. Clark, and P. Green, “Synthesized Clustering: A Method for Amalgamating Clustering

Bases with Differential Weighting Variables, "Psychometrika, vol. 49, no. 1, pp. 57-78, 1984.

[20] P. Green, J. Kim, and F. Carmone, "A Preliminary Study of Optimal Variable Weighting in K-Means Clustering," J. Classification, vol. 7, no. 2, pp. 271-285, 1990.