

Wine Tasting and a Novel Approach to Cluster Analysis

William Cluster
Asia Pacific Management
Ritsumeikan Asia Pacific University
Beppu, Japan
wcluster@apu.ac.jp

Subana Shanmuganathan and Philip Sallis
Geo-Informatics Research Center
Auckland University of Technology
Auckland, New Zealand
Sshanmug@aut.ac.nz

Abstract — This paper proposes an alternate view for understanding clusters and for determination of the variables that play a significant role in cluster makeup. We explore the creation of a new categorical variable from a given set of variables by means of the output of a clustering algorithm. We postulate that this new variable can be seen as being comprised of a few “important” variables and explore how this new variable relates to the original variables. Sensitivity analysis and discriminant analysis are used to confirm the selection of important variables. Then we show that this method is able to identify those key variables which vary most significantly throughout the clusters.

Key words: Clustering, Variable selection after clustering, Multivariate analysis, Distance-based clustering, TwoStep Clustering, K-Means, SOM, binning, discriminant analysis, Sensitivity analysis.

I. INTRODUCTION

A. Background

The initial goal of cluster analysis is to partition a data set of N objects into subgroups such that those in each particular group are more similar to each other than to those of other groups. This is done by inputting the variables into a clustering algorithm [1].

The results of this process are generally used to segment data which is relatively little understood in order to perceive some structure in the data. Often, in a kind of divide and conquer strategy, after the data has been clustered, the analyst will focus on individual clusters in order to understand their makeup, taking into account how the input variables vary over the clusters; attempting to ascertain in which clusters the mean of an input variable is high and in which it is low, and further; in which cluster the dispersion of an input variable is high and in which cluster it is low. This leads to an understanding

of the clusters and thereby a plausible understanding of the structure of the data [2] [3] [4] [5] [6] [7].

Here we will suggest a method for identifying important variables. We also show that this can lead to a richer understanding of individual clusters because these are among those variables which vary most from cluster to cluster within the clustering. In the process we suggest an approach to measuring proximity of variables. This measure of proximity is somewhat simpler than correlation.

Note: In this discussion we will often need to refer to the “number of values” that a categorical variable can take on. We will henceforth refer to this as the “cardinality of the measurement scale”. Also note: If the number of clusters does not match the cardinality of the measurement scale of all of the variables, our technique breaks down, and so we choose to set the number of clusters to n and also to “bin” each variable into the same n possible values. The method for binning is described below.

II. DATA DETAILS

We first analyzed data which was mostly binary (0/1) but that also had categorical variables. We sought a high dimensional data set. The data set used arises from a text mining experiment on wine tastings. The data consists of wines of different styles and from different regions within New Zealand. Taster comments of these wines were mined using standard text mining techniques [8]. The text mining procedure resulted in a data set consisting of 61 variables recorded for each of 95 different wines. Thus each wine was measured on 61 variables. For categorical variables with more than 2 values we binned them into 2 bins using a K-Means algorithm as explained in the section below labeled “Binning Categorical Variables Using a Clustering Algorithm”.

The data set being binary data suggests that this method is applicable in many research settings as binary outcomes arise frequently [9].

III. PROXIMITY MEASURE

Data is generally collected on a set of instances (also known as records, individuals, cases, observations, etc.) for generally a fairly small number of variables. Nonetheless each record could conceivably be measured with respect to a huge number of variables. In this paper we view a clustering of the data set as a new variable and investigate circumstances under which this new variable may be related to the original input variables. This approach differs from the usual interest in a clustering exploration. Usually in a clustering exploration the individual clusters are the focus and the nature of each individual cluster is investigated..

Here we view the clustering itself as a new categorical variable and try to understand how it may be related to the original input variables. We develop a means of identifying variables that are related to the overall cluster pattern (the new categorical variable) and use this method to identify which variables in the original data are most closely related to the overall clustering. We will refer to this method as our “proximity measure” from here on.

As a means of comparison and confirmation, we use discriminant analysis where the grouping variable was our new cluster variable and all the input variables were the independent. We used the discriminant analysis algorithm provided by SPSS. We found that there is significant agreement our proximity measure and the discriminant analysis.

We also apply a sensitivity analysis to the same problem and again find significant agreement in terms of identification of the original variables which are most closely related to the cluster pattern.

Then, most importantly, we show that the variables ranking highest in our proximity measure method are among those that rank highest in terms of those variables that vary most over the clusters. Thus these are among the most important to the cluster analysis.

IV. DATA SET: WINE TASTINGS

Our data came from a set of verbal descriptions of wines from various regions within New Zealand and of various types. Our data was processed for text mining using a stemming algorithm and resulted in a total of 61 variables; each one being a stemmed piece of text. (see table 1). We then used a clustering algorithm to cluster the data. We choose the SPSS TwoStep Clustering algorithm for this purpose because it is fairly flexible in the data types it will accept; accepting both continuous and categorical variables.

For other data sets it may be of interest to try SOM, K-Means, or Hierarchical Clustering. Note that this kind of text mining results in sparse, 0/1 vectors, where each record will consist primarily of zeros and

there will only be a 1 when that “word” was used to describe the particular wine designated by that record.

TABLE I. LIST OF 60 VARIABLES OBTAINED FROM THE TEXT COMMENTS OF 95 WINES.

wtype	Region	price	year	rate
ag	appl	apricot	barrel	berri
black	blend	bodi	cellar	chardonnai
cherri	chocol	cinnamon	citru	citrusi
cola	creami	dusti	ferment	fig
floral	fresh	gooseberri	grapefruit	herb
herbal	honei	lime	melon	miner
nectarin	oak	orang	passion	peach
pepper	pineappl	pinot	plum	red
riesl	sauvignon	silki	slightli	smoke
smoki	soft	spice	sweet	tannin
toast	toasti	tropic	vanilla	young

Typically in such a study, after generating a clustering of the data, the analyst would focus on the individual clusters to see if knowledge can be gathered from understanding these clusters individually. However our aim is to view the group of clusters as a whole and try to understand them as possibly representing a single concept. It will be convenient to use the term ‘clustering’ as a noun to refer to the partition of the entire set of data and then to use the term ‘cluster’ to refer to an individual subgroup of data points that is formed by the clustering and such that these data points are deemed similar by the clustering algorithm. (See figure 1.) Thus, instead of focusing on the individual clusters we treat the clustering as a new variable and try to understand as a single unit.

In figure 1 below, we see an illustration of a cluster variable defined by the large black X which divides the data set up into BL-C1 (black lines, cluster 1), BL-C2, BL-C3, and BL-C4. The data set is the entire rectangular region. Overlaid on this is another variable we call R. The dividing lines of R are in orange (grey if the picture is not in color) not black. R is a categorical variable and thus separates the data set into subsets whose points assume the same values of R throughout each of these subsets. The figure illustrates both equation 1 and equation 2.

$$BLC4 \subset RC3 \quad (1)$$

$$BLC3 \subset RC3 \cup RC4 \cup RC2 \quad (2)$$

Note that equation (2) illustrates that a cluster may have more than one feature of a variable.



Figure 1. This is a visual representation of the data set. Here BL is the clustering variable and R represents some other arbitrary categorical variable. The figure illustrates that any of the clusters in the BL clustering is a contained in the union of some values of R. The clusters R-C_i are illustrated with different colors whereas the partition due to the variable BL are shaded in grey.

A. Description of Methods

In the discussion below we will use the term ‘input variable’ to mean any variable that was not created by the clustering process. We will also use the term ‘clustering’ and ‘cluster variable’ synonymously since we are thinking of the clustering as creating a new variable.

Several clustering algorithms were considered including TwoStep Clustering, K-Means, Hierarchical Clustering, and Self Organizing Maps. The data was both numeric and categorical and so a TwoStep Clustering algorithm available through SPSS version 16 was selected. In terms of the settings in SPSS, a Log-likelihood distance measure was selected, no special outlier procedure was used, and most importantly, a cluster membership variable was created. We wanted to see if any of the measured variables approximated the clustering and if so how closely.

In attempting to know which, if any, measured variables the clustering may represent, we needed to consider how to calculate the distance between two variables. That is, we had 61 input variables and 1 cluster variable, and we wondered which of the input variables was “closest” to the cluster variable. We adopted a simple measure of distance between variables as described next.

B. Measure of Distance between Two Variables

The distance between two variables was measured by comparing those variables for each record in our data set. For each record where the two variables differed we added 1 to the calculation of the total distance between these two variables. For example, let V₁ be one of the measured variables and C be the cluster variable. Similar to the concept of a random variable in probability theory we will consider these variables to be functions from the data set to the set of

values that they can take on and we will label the values they can take on as v_j and c_j respectively. In the first case, the data set is the set of wines and so we will label each element of the data set w_i. Furthermore let W be the set of all records in the data set (in other words; all the wines). Thus we can write V₁(w_i) = v_j and C(w_i) = c_j. For each w_i in W we compared V₁(w_i) and C(w_i). Whenever they were not the same we added a 1 to the calculation of the distance between V₁ and C.

Distance between V₁ and C:

$$d(V_1, C) = \sum_j \delta_{V_1(w_j) C(w_j)} \quad (3)$$

where δ is the Kronecker delta function being 0 when the arguments are the same and 1 when they are different and where the summation is taken over all the wines w_j in W.

With this as our measure of distance we looked for those variables which were closest to the cluster variable. For example the least distance would be

$$\text{Min}_{V_i} [d(V_i, C)].$$

After calculating the distances of each input variable V_i to the cluster variable C we ordered all the V_i by their distance from C.

C. Additional Issue of Permutations

There was an additional consideration with regard to finding the distance of V_i variable to C. A clustering C is a partition of the data set. Our question was: how close is this partitioning to a given measured categorical variable, say V₁?

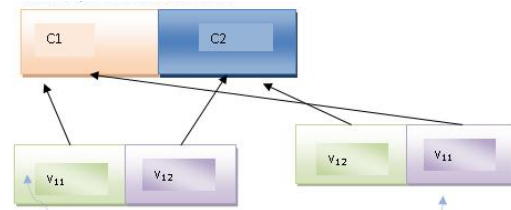


Figure 2. Illustration of 2! (=2) ways of mapping the values of V₁; v₁₁ and v₁₂, to the clusters C₁ and C₂ of the clustering C. Note that the data set is partitioned by C (above) into C₁ and C₂ or partitioned by V₁ (either of the rectangles below) into v₁₁ and v₁₂.

For example, suppose that C breaks the data set into two subsets which we can denote by C₁ and C₂ and that V₁ takes on the two values v₁₁ and v₁₂. We are using the notation v_{ij} where i is the index indicating which variable is being considered and j is the index indicating which value of that variable is being designated.

As it is the goal of this paper to suggest that perhaps V₁ and C are either the same or similar

variables (they represent the same partitioning of the data set) we have sought to find out over how many records they are the same. It would of course be optimal if for every record they concur. This would mean that for each wine w_k , $V1(w_k) = C(w_k)$. More specifically, when, for each k where w_k resulted in the value v_{11} then for those same k , w_k would result in the value $c1$ for the cluster variable C . But then we recognize that since the meaning of these clusters is unknown there is no particular significance to v_{11} always being associated with $c1$ and v_{12} being associated with $c2$. It would be equally as notable if v_{11} were consistently associated with $c2$ and v_{12} were consistently associated with $c1$. In the current case, where both the variable and the clusters only partition the data set into 2 subsets this leaves us with $2! = 2$ ways of matching the clustering with our variable $V1$ (see figure 2). In general there would be $n!$ ways of matching the clustering with our variable V_i , where n is the number of clusters in the clustering. We opt to choose the matching that gives us the least distance from the variable to the cluster (where distance is defined in the previous discussion). It is possible that this is not the true matching but since we have no way of knowing this we choose the one with a minimum distance. Our strategy was to consider all $n!$ possibilities and then choose the permutation that resulted in the minimum distance between the two variables. After this minimum was established then the same procedure would be run for all of the measured variables $V2$, $V3$, and so on up to the last variable.

This procedure can be expressed as seeking the permutation of the values of C which minimizes the distance with V_i :

$$\text{Min}_{\forall m} [d(V_i, C_m)] \quad (4)$$

where m runs through all $n!$ permutations of the values of C .

Then as described earlier (in the section: Measure of Distance between Two Variables) the variables would be ranked according to their distances from the cluster variable.

In addition to the wine data, we attempted the same procedure on a second data set where we considered a clustering with 7 clusters. In that case there were input variables that took on as many as 16 values.

D. Binning Categorical Variables Using a Clustering Algorithm

Since there is no standard method to bin categorical variables we used the following strategy. Again we used the SPSS TwoStep algorithm with all the categorical variables as inputs. We forced the number of clusters to be 2 for our wine data set (and 7 for our optical character recognition data set). The

algorithm then assigns each case to a cluster. Since one value of a categorical variable may be assigned to more than one cluster in this process we choose the most frequently assigned cluster in which to bin the variable. In the case of a tie a random selection was made within those tied clusters. Note that there are other choices rather than the TwoStep algorithm for clustering categorical variables [10] [11] [12] [13] [14].

E. Exploration and Confirmation

From the list of input variables (see Table 1) we used the TwoStep clustering algorithm supplied by SPSS to create a new cluster variable TSC_2 (our naming comes from Two Step Clustering generating 2 clusters). We generated a clustering with 2 clusters. We were interested to see if the cluster variable TSC_2 might represent any of the input variables, i.e. did TSC_2 approximate any of these variables or alternatively which of the variables did it approximate most closely.

Using the above procedures we were able to come up with the following ranking of input variables; from closest to the cluster variable to farthest from the cluster variable.

Table 2 suggests that if TSC_2 is a replica of any of the variables then the best candidates, ranked in order, may be 'cherri', 'black', 'vanilla'... etc. Furthermore with regard to the significance of the results, using either a binomial or chi square calculation of the probability of getting an error of 26 or less gives 0.000006 for binomial or $p < 0.001$ for chi squared ($\chi^2 = 24$, $df = 1$), suggesting that the match between cherri and TSC_2 is not a random match.

Next, we seek to corroborate this procedure by other means and thereby confirm that the variables with least distance to the cluster variable TSC_2 might in fact be important in relation to TSC_2 . We used both discriminant analysis and then also sensitivity analysis to ascertain which variables may be important in predicting cluster membership. Both approaches produced a similar list of "important" variables as described next.

F. Discriminant Analysis.

A discriminant analysis is conducted with SPSS version 16 to confirm the results [15]. Using the Wilks' Lambda test statistic to ascertain which variables were strong predictors of TSC_2 . There is considerable consistency of results between our proximity measure method and discriminant analysis (see table 3). 8 of the top 10 variables that were closest to being replicated by the clustering variable were also in the top 10 list of important variables in the discriminant analysis.

In particular, the following variables were included in the top 10 of each list: cherri, black, vanilla, pinot, cinnamon, tannin, plum, chocol. Using the 'N - 1' chi-squared test (the K. Pearson chi-squared

test but with N replaced by N - 1)¹⁶ we get 'N-1' Chi squared = 34.08, p < 0.0001 strongly suggesting that this agreement is not random.

Although the selection of these variables was arrived at with seemingly unrelated methods they both produce highly similar results. The results may suggest that the cluster variable TSC_2 is in some sense a replica of one or more of the highest ranked variables from Table 2.

G. Sensitivity Analysis

We also ran a sensitivity analysis of the importance of the measured variables in predicting the cluster variable. In the data mining package Clementine, when creating a neural network, one of the reports that is produced is a ranking of the variables using sensitivity analysis. Table 4 shows the results of this analysis.

Although the statistical significance was not as high as for the discriminant analysis, within the top 10 ranked in the sensitivity analysis, 3 (black, soft, herb) were also ranked within the top 10 by our process. Again using 'N - 1' chi-squared test we get 'N-1' Chi squared = 1.91, P = 0.2

V. RESULTS

After confirming our proximity measure using discriminant analysis and sensitivity analysis we sought to discover whether the notion of proximity to the cluster variable can be useful in cluster analysis. As stated earlier a primary application of cluster analysis is in understanding the nature of the individual clusters. With the clustered wine data we checked to see which variables varied most over the clusters. We computed the means, standardized means, and standard deviations for each variable in each cluster. As there were only two clusters for this data set, we computed the difference in the means of cluster 1 and cluster 2 for each variable. We were interested in which variables had the greatest difference with respect to these two clusters. We hoped that our proximity measure methodology of choosing important variables would also single out those variables that varied the most over the two clusters. (See table 5 below).

Within the top 10 variables that had greatest variation in means, our proximity measure had found 5. Again using 'N-1' Chi squared test we found that 'N-1' Chi squared = 14.06, with P = 0.002.

We did the same test for the standardized means and the standard deviation. For the standardized means, from the top 10 variables selected by our method, 4 were within the list of greatest variation¹

¹ There were a number of variables where the standard deviation was zero within one of the clusters. In that case we set the standardized value to zero.

and thus 'N-1' Chi squared = 4.94 with P = 0.03. Within the top 10 variables that had greatest variation in standard deviation, the proximity measure had found 5. Again using 'N-1' Chi squared test we found that 'N-1' Chi squared = 14.06, with P = 0.002.

VI. DISCUSSION

Trying to understand the clustering process and the results it produces leads to insights into original dataset. A common approach to analyzing clustering is to focus on the individual clusters. Here our aim has been to understand the nature of the clustering as a whole and in particular to see it as some sort of summation of seen and perhaps unseen variables. In this paper we described how a measure of distance between variables can be used to identify those variables of which the clustering may be a replica. We have suggested that these variables may be key to understanding the nature of the individual clusters and thereby assist an analyst in gaining knowledge about these individual clusters. We have confirmed that the proximity measure described here is successful in identifying key variables in the clustering.

TABLE II. RANKING OF THE INPUT VARIABLES IN TERMS OF THEIR PROXIMITY TO THE CLUSTER VARIABLE TSC_2.

Variable	distance	Variable	distance
cherri	26	floral	34
black	26	red	35
vanilla	32	dusti	36
pinot	32	cola	36
cinnamon	32	cellar	36
tannin	33	bodi	36
soft	33	herbal	37
plum	33	pepper	39
herb	33	spice	40
chocol	33	slightli	40
silki	34	berri	40

TABLE III. VARIABLES AND THEIR WILK'S LAMBDA COEFFICIENT SHOWING RANKING OF STRENGTH OF PREDICTOR OF TSC_2

black	0.8	pineappl	0.9
cherri	0.8	nectarin	0.9
cinnamon	0.9	herb	0.9
pinot	0.9	soft	0.9
appl	0.9	citru	0.9
vanilla	0.9	honei	0.9
chocol	0.9	miner	0.9
cola	0.9	oak	0.9
plum	0.9	toasti	0.9
tannin	0.9	ferment	0.9
peach	0.9	gooseberri	0.9
lime	0.9	melon	0.9

Significance level for 29 highest ranked variables each less than .05

TABLE IV. SENSITIVITY ANALYSIS FROM SPSS CLEMENTINE NEURAL NETWORK NODE.

1	black	13	nectarin
2	soft	14	vanilla
3	herb	15	sweet
4	fig	16	appl
5	red	17	cherri
6	smoke	18	citrusi
7	lime	19	peach
8	gooseberri	20	herbal
9	young	21	chocol
10	floral	22	ferment
11	tannin	23	oak
12	berri	24	grapefruit

Variables ranked per importance in predicting TSC_2

TABLE V. LISTING OF IMPORTANT VARIABLES FOR DISTINGUISHING AMONG CLUSTERS.

Ranking of 10 Closest Variables to TSC_2	Ranking of Largest Variation between Clusters in Non-standardized Means	Ranking of Largest Variation between Clusters in Standardized Means	Ranking of Largest Variation between Clusters in Standard Deviation
cherri	Black ^a	black	black
black	cherri	cherri	cherri
vanilla	peach	appl	appl
pinot	appl	lime	lime
cinnamon	lime	pineappl	pineappl
tannin	pineappl	cinnamon	cinnamon
soft	vanilla	pinot	pinot
plum	cinnamon	nectarin	nectarin
herb	pinot	citru	chocol
chocol	nectarin	honei	cola

^a Yellow shading highlights variables identified by proximity measure that are among the top 10 ranked variables.

[1] J. H. Friedman, J. J. Meulman . Clustering Objects on Subsets of Attributes , Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 66, No. 4 (2004), pp. 815-849.

[2] G. Punj and D. W. Stewart, Cluster Analysis in Marketing Research: Review and Suggestions for Application. Journal of Marketing Research, Vol. 20, No. 2 (May, 1983), pp. 134-148.

[3] I. Bose, X. Chen, Quantitative models for direct marketing: A review from systems perspective, European Journal of Operational Research, Volume 195, Issue 1, 16 May 2009, Pages 1-16, ISSN 0377-2217, DOI: 10.1016/j.ejor.2008.04.006.

[4] J. Bowen. Development of a taxonomy of services to gain strategic marketing insights. Journal of the Academy of Marketing Science. Volume 18, Number 1 / December, 1990. Pages 43-49.

[5] C.-Y. Chiu, Yi.-F. Chen, I.T. Kuo, H. C.Ku, An intelligent market segmentation system using k-means and particle swarm optimization, Expert Systems with Applications, Volume 36, Issue 3, Part 1, April 2009, Pages 4558-4565.

[6] M. Wedel, W. A. Kamakura Market Segmentation: Conceptual and Methodological Foundations, Edition: 2, ISBN 0792386353, 9780792386353.

[7] S. Zani, A. Cerioli, M. Riani, M. Vichi Contributor Sergio Zani Data Analysis, Classification and the Forward Search: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6-8, 2005 By Classification Group of SIS. Meeting, page 23, ISBN 354035977X, 9783540359777.

[8] P.J. Sallis, S. Shanmuganathan, L. Pavesi and M.C. Jarur Muñoz. Kohonen Self-organising maps in the mining data mining of wine taster comments. Data Mining IX, Data Mining, Protection, Detection and other Security Technologies 2008. Cadiz, Spain, 26-28 May 2008. Eds., A Zanasi, D Almorza Gomar, N F F Ebecken and C A Brebbia. ISBN: 978-1-84564-110-8, ISSN (print): 1746-4463, ISSN (on-line): 1743-3517 Transactions on information and Communication Technologies, Vol. WIT press. 40 pp 125-139.

[9] Within-Cluster Resampling E. B. Hoffman, P. K. Sen, C. R. Weinberg Biometrika, Vol. 88, No. 4 (Dec., 2001), pp. 1121-1134 Published by: Biometrika Trust , Stable URL: <http://www.jstor.org/stable/2673705>.

[10] R. Baragona, C. Calzini, F. Battaglia, (2001) Genetic algorithms and clustering: an application to Fisher's Iris Data, in S. Borra et al., eds., Advances in Classification and Data Analysis, Springer-Verlag, Berlin/Heidelberg.

[11] Z. Huang (1998) A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In: Proc. ACM SIGMOD Workshop on Data Mining and Knowledge Discovery, pp. 146-151.

[12] L. Kauffman, P.J. Rousseeuw (1990) Finding Groups in Data, Wiley & Sons, NY.

[13] A fuzzy k-modes algorithm for clustering categorical data Zhexue Huang M.K. Ng, Manage. Inf. Principles Ltd., Melbourne, Vic.; This paper appears in: Fuzzy Systems, IEEE Transactions on Publication Date: Aug 1999 Volume: 7, Issue: 4 On page(s): 446-452 ISSN: 1063-6706 References Cited: 18 CODEN: IEFSEV INSPEC Accession Number: 6350896 Digital Object Identifier: 10.1109/91.784206 Current Version Published: 2002-08-06.

[14] <http://www.statsoft.com/textbook/stcluan.html>.

[15] R. Khattree and D.N. Naik, 2000, Multivariate data reduction and discrimination, SAS Institute, Cary, North Carolina.

[16] Campbell, I. (2007) Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. Statistics in Medicine, 26, 3661-3675.