Empirical Comparative Analysis of 1-of-K Coding and K-Prototypes in Categorical Clustering

Fei Wang[1], Hector Franco[1], John Pugh[2], and Robert Ross[1]

1 School of Computing, Dublin Institute of Technology, Ireland
2 Nathean Technologies Ltd. Dublin, Ireland.

d13122837@mydit.ie

Introduction

Clustering is a fundamental machine learning application, which partitions data into homogeneous groups. K-means and its variants are the most widely used class of clustering algorithms today. In this paper, we compare two methods, 1-of-K coding and k-prototypes, in categorical data clustering.

Background

Our own interest in clustering stems from its importance in customer segmentation. We are particularly concerned with the data with a high proportion of categorical data, as it is the most common form of customer data.

1-of-K Coding and K-prototypes

- The common method 1-of-K coding converts each category feature into a set of binary features using 1 and 0 to represent a category value present or absent in objects;
- K-prototype on the other hand inherits the ideas of k-means, but applies the simple matching distance and modes to categorical features.

Results

From *Fig. 1* and *Fig. 2*, it is shown that when the dataset gets large, the time consumed for k-means with 1-of-K coding is 2 to 3 times greater than that for k-prototypes.

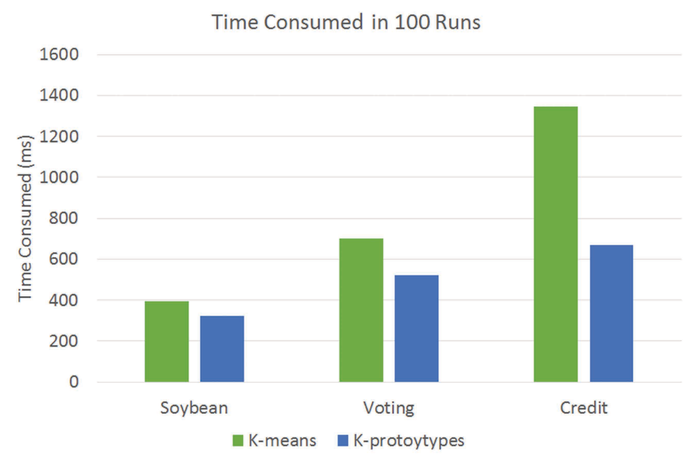

Fig.1: Time Consumed - Soybean, Voting and Credit

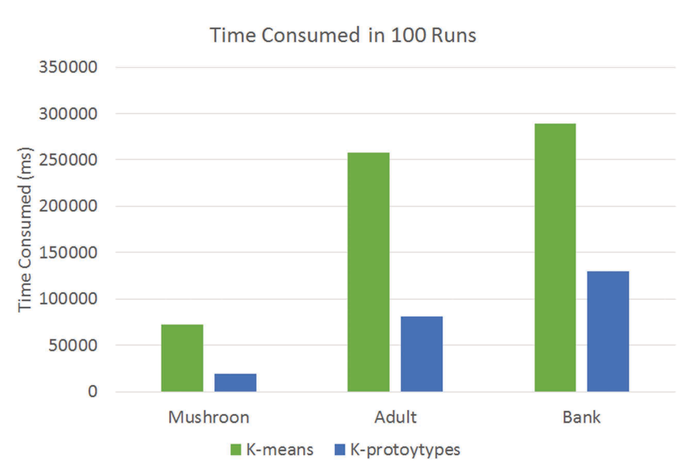

Fig.2: Time Consumed - Mushroom, Adult and Bank

However, from *Fig. 3* and *Fig. 4*, we can see that the k-means algorithm consumes much more time not because it needs more iterations to converge, but because 1-of-K coding substantially expands the dimensionality of the datasets.

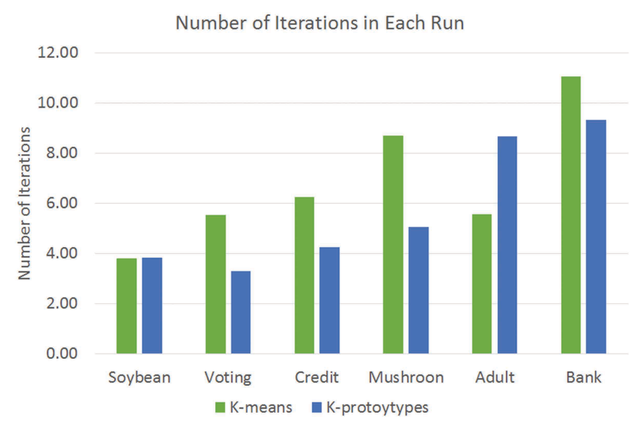

Fig.3: Number of Iterations in Each Run

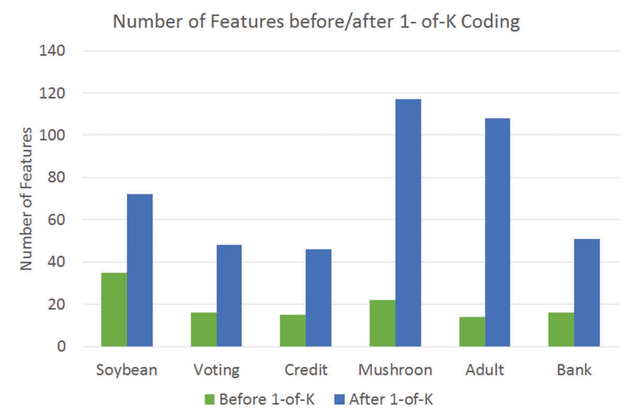

Fig.4: Number of Features before/after 1-of-K Coding

Fig. 5, *Fig. 6* and *Fig. 7* summarise the accuracy calculation results of three datasets that have the accuracies correlated with the cost functions.

Accuracy Table - Soybean		
	Kmeans	Kprototypes
100%	**46**	**22.13**
99%-100%	0	0.00
98%-99%	0	0.00
97%-98%	0	**8.47**
96%-97%	0	0.00
95%-96%	0	**12.27**
94%-95%	0	0.00
93%-94%	0	0.00
92%-93%	0	0.00
91%-92%	0	2.27
90%-91%	0	0.00
<90%	54	54.86

Fig. 5: Accuracy Table - Soybean

Accuracy Table - Voting		
	Kmeans	Kprototypes
88%-89%	**96**	0.00
87%-88%	0	0.00
86%-87%	0	**75.33**
85%-86%	0	**24.67**
84%-85%	0	0.00
83%-84%	0	0.00
82%-83%	0	0.00
81%-82%	0	0.00
80%-81%	0	0.00
79%-80%	0	0.00
78%-79%	0	0.00
<78%	4	0.00

Fig.6: Accuracy Table - Voting

1. Both algorithms get almost the same highest accuracy. The differences are only 1% - 2%;
2. The valid results with k-means concentrate at the interval of highest accuracy, while the ones with k-prototypes spread much more widely in the valid range, that is, k-means is more stable than k-prototypes;

Accuracy Table - Mushroom		
	Kmeans	K-prototypes
89%-90%	**57**	**10.80**
88%-89%	0	14.20
87%-88%	0	1.67
86%-87%	0	1.47
85%-86%	0	5.27
84%-85%	0	0.00
83%-84%	0	0.00
82%-83%	0	1.00
81%-82%	0	1.20
80%-81%	0	0.93
79%-80%	0	7.60
<79%	43	55.86

Fig.7: Accuracy Table - Mushroom

3. The results in bold refer to the objectively best results based on cost function. K-means probably finds only one global optimum, but k-prototypes can find multiple global optima.

Conclusion

Even though they use different distances in calculating dissimilarity, k-means with 1-of-K coding and k-prototypes provide similar best results. For the clustering speed, k-prototypes is faster than k-means with 1-of-K coding, because the latter expands significantly the dimensionality of the dataset. For the clustering validity, the valid results with k-prototypes spread in multiple optima, while the ones with k-means with 1-of-K coding concentrate in one point. Therefore, we conclude that k-means with 1-of-K coding is more stable than k-prototypes.