Robust Qualitative Data Clustering via Learnable Metric Space Fusion

Sen Feng[1], Mingjie Zhao[1], Zhanpei Huang[1], Yuzhu Ji[1], Yiqun Zhang[1,2,*], Yiu-Ming Cheung[2]
[1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[2]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

Proof. Assume δ_1 is the average number of iterations per weight update and δ_2 represents the additional iterations required for the convergence of **Q'**.

For each weight update, the distances between N objects and k modes across M metrics are calculated. Since the average number of iterations is δ_1, and the time complexity of searching the distance between two samples is $O(d)$, the time complexity for this phase is $O(M\delta_1 Nkd)$. During the fusion phase, each update of **Q'** is accompanied by a re-evaluation of the distances between each sample, the time complexity for this phase is $O(M\delta_2 Nkd)$.

In summary, assume that the weights are updated l times in total, the overall time complexity is $O(Ml(\delta_1 + \delta_2)Nkd)$. Therefore, the time complexity can be simplified to $O(M\delta lNkd)$, where δ represents the sum of constant terms δ_1 and δ_2.

Furthermore, in terms of the required storage space for the algorithm, we only need to store the distances between all points and modes under each metric. Therefore, the time complexity can be summarized as $O(MNk)$. □

Since static data only needs to be calculated once for different metrics, in order to simplify the analysis process, the theorem does not analyze the complexity of calculating the similarity matrix between possible values based on the base metric. If readers need to know more, please read the original paper of the metric we use.

I. Complementary Experimental Results

Clustering performance of different methods and significance test are compared in Table I and Table II. The bast and second-bast results are marked in orange and grey on each dataset, respectively. − indicates that this method fails to run on this dataset. Significance test is conducted between the best and the second-best results by Wilcoxon signed rank test [**?**] with a 95% confidence interval, and significant difference is indicated by the symbol " • ". The observations include the following ... aspects: (1) MSF method achieves the best performance in 6 out of 10 datasets based on CA, while it achieved the second-best result in the LG dataset. (2) For some imbalanced datasets such as CE, LE, and LG, the SCPP-a method tends to assign a large number of samples to certain large clusters, while small clusters that should exist are occupied by only a few samples. As a result, due to the way the CA metric is calculated, SCPP-a performs better on imbalanced datasets. However, its performance in terms of NMI shows a significant decrease. (3) The MSF method achieves the best performance in 5 out of 10 datasets based on the CA metric, while it achieved the second-best result in the CE, LG, AC datasets.

* Corresponding author: Yiqun Zhang (yqzhang@gdut.edu.cn)

TABLE I

CLUSTERING PERFORMANCE EVALUATED BY CA. ORANGE AND GRAY REPRESENT THE BEST AND SECOND-BEST RESULT. THE SYMBOL "•" INDICATES THAT THE MSF METHOD AND OTHER COMPARATIVE METHODS HAVE PASSED THE 95% CONFIDENCE TEST BASED ON THE WILCOXON SIGNED-RANK TEST.

Dataset	ECPCS-HC^I	ECPCS-HC	LWEA	ELSC	LROLML-2	SCPP-a	H2H	COForest	MSF(ours)
TT	0.5294±0.03	0.6162±0.00	0.6204±0.02	0.6208±0.00	0.6406±0.04	0.6526±0.00	0.5433±0.00	0.5731±0.00	0.6816±0.00 •
CS	0.5925±0.06	0.5150±0.01	0.5138±0.01	0.6013±0.00	0.6038±0.04	0.5863±0.04	0.6175±0.02	0.6313±0.04	0.6500±0.00
CE	0.4108±0.03	0.5946±0.04	0.5937±0.05	0.3983±0.05	-	0.6987±0.00	0.3955±0.11	0.3542±0.04	0.5883±0.00
VT	0.8713±0.00	0.8437±0.01	0.8600±0.03	0.8391±0.00	0.7224±0.07	0.8626±0.00	0.8736±0.00	0.8761±0.00	0.8782±0.00 •
LG	0.4730±0.04	0.4966±0.03	0.4956±0.03	0.4770±0.01	0.4628±0.08	0.5662±0.05	0.5250±0.04	0.5088±0.07	0.5608±0.00
AC	0.7942±0.00	0.5106±0.01	0.5330±0.05	-	-	0.6981±0.12	0.7703±0.08	0.7857±0.11	0.8087±0.00 •
ER	0.1986±0.01	0.1943±0.01	0.1917±0.01	0.1723±0.01	0.2010±0.01	0.1948±0.01	0.1836±0.01	0.1929±0.01	0.1920±0.00
DT	0.7492±0.06	0.5492±0.08	0.5455±0.06	0.7472±0.00	0.5075±0.08	-	0.7800±0.04	0.7048±0.08	0.7877±0.00
AE	0.5681±0.06	0.6000±0.04	0.5889±0.04	0.5042±0.05	0.5278±0.05	0.4292±0.07	0.5181±0.07	0.5583±0.09	0.6944±0.00 •
LE	0.3116±0.02	0.3614±0.02	0.3504±0.02	0.3014±0.00	0.3526±0.03	0.4043±0.00	0.3144±0.02	0.3155±0.2	0.3240±0.00

TABLE II

CLUSTERING PERFORMANCE EVALUATED BY NMI. ORANGE AND GRAY REPRESENT THE BEST AND SECOND-BEST RESULT. THE SYMBOL "•" INDICATES THAT THE MSF METHOD AND OTHER COMPARATIVE METHODS HAVE PASSED THE 95% CONFIDENCE TEST BASED ON THE WILCOXON SIGNED-RANK TEST.

Dataset	ECPCS-HC^I	ECPCS-HC	LWEA	ELSC	LROLML-2	SCPP-a	H2H	COForest	MSF(ours)
TT	0.0071±0.00	0.0071±0.00	0.0070±0.10	0.0420±0.00	0.0041±0.01	0.0016±0.00	0.0110±0.01	0.0106±0.01	0.1052±0.00 •
CS	0.0536±0.04	0.0204±0.00	0.0265±0.01	0.0287±0.00	0.0475±0.03	0.0780±0.04	0.0448±0.02	0.0749±0.03	0.0975±0.00 •
CE	0.1731±0.05	0.0402±0.01	0.0407±0.02	0.0349±0.01	-	0.0046±0.01	0.0715±0.05	0.1132±0.07	0.1163±0.00
VT	0.4901±0.01	0.4073±0.04	0.4355±0.06	0.4409±0.00	0.1667±0.11	0.4515±0.00	0.4893±0.00	0.4897±0.00	0.4993±0.00
LG	0.1328±0.04	0.1546±0.02	0.1701±0.03	0.2392±0.01	0.1088±0.04	0.1594±0.07	0.2172±0.02	0.1811±0.05	0.2311±0.00
AC	0.2668±0.00	0.0451±0.00	0.0537±0.04	-	-	0.1615±0.13	0.2402±0.08	0.3042±0.17	0.3027±0.00
ER	0.0667±0.00	0.0708±0.00	0.0680±0.00	0.0355±0.00	0.0614±0.01	0.0411±0.02	0.0494±0.00	0.0049±0.01	0.0533±0.00
DT	0.7948±0.05	0.6367±0.04	0.6196±0.03	0.7171±0.00	0.4501±0.10	-	0.8023±0.02	0.8089±0.06	0.8567±0.00 •
AE	0.1934±0.05	0.2076±0.07	0.2112±0.05	0.1892±0.05	0.2312±0.07	0.0816±0.09	0.2249±0.08	0.2340±0.08	0.3795±0.00 •
LE	0.0551±0.01	0.0467±0.02	0.0319±0.01	0.0375±0.00	0.0638±0.01	0.0080±0.00	0.0574±0.02	0.0560±0.02	0.0573±0.00