

Contents

1	Cluster Analysis and K-means Clustering: An Introduction	1
1.1	The Emergence of Data Mining	1
1.2	Cluster Analysis: A Brief Overview	2
1.2.1	Clustering Algorithms	3
1.2.2	Cluster Validity	5
1.3	K-means Clustering: An Ageless Algorithm	7
1.3.1	Theoretical Research on K-means	8
1.3.2	Data-Driven Research on K-means	9
1.3.3	Discussions	11
1.4	Concluding Remarks	12
	References	12
2	The Uniform Effect of K-means Clustering	17
2.1	Introduction	17
2.2	The Uniform Effect of K-means Clustering	18
2.2.1	Case I: Two Clusters	18
2.2.2	Case II: Multiple Clusters	20
2.3	The Relationship Between K-means Clustering and the Entropy Measure	23
2.3.1	The Entropy Measure	23
2.3.2	The Coefficient of Variation Measure	23
2.3.3	The Limitation of the Entropy Measure	24
2.4	Experimental Results	25
2.4.1	Experimental Setup	25
2.4.2	The Evidence of the Uniform Effect of K-means	27
2.4.3	The Quantitative Analysis of the Uniform Effect	28
2.4.4	The Evidence of the Biased Effect of the Entropy Measure	30
2.4.5	The Hazard of the Biased Effect	31

2.5	Related Work	33
2.6	Concluding Remarks	34
	References	34
3	Generalizing Distance Functions for Fuzzy c-Means Clustering . .	37
3.1	Introduction	37
3.2	Preliminaries and Problem Definition.	39
3.2.1	Math Notations.	39
3.2.2	Zangwill’s Global Convergence Theorem	39
3.2.3	Fuzzy c-Means	40
3.2.4	Problem Definition	42
3.3	The Point-to-Centroid Distance	43
3.3.1	Deriving the Point-to-Centroid Distance.	43
3.3.2	Categorizing the Point-to-Centroid Distance.	48
3.3.3	Properties of the Point-to-Centroid Distance.	48
3.4	The Global Convergence of GD-FCM	49
3.5	Examples of the Point-to-Centroid Distance	54
3.6	Experimental Results	56
3.6.1	Experimental Setup	56
3.6.2	The Global Convergence of GD-FCM	61
3.6.3	The Merit of GD-FCM in Providing Diversified Distances.	61
3.7	Related Work	63
3.8	Concluding Remarks	64
	References	65
4	Information-Theoretic K-means for Text Clustering	69
4.1	Introduction	69
4.2	Theoretical Overviews of Info-Kmeans	70
4.2.1	The Objective of Info-Kmeans	71
4.2.2	A Probabilistic View of Info-Kmeans	71
4.2.3	An Information-Theoretic View of Info-Kmeans.	72
4.2.4	Discussions	74
4.3	The Dilemma of Info-Kmeans.	74
4.4	The SAIL Algorithm	75
4.4.1	SAIL: Theoretical Foundation	76
4.4.2	SAIL: Computational Issues.	77
4.4.3	SAIL: Algorithmic Details	80
4.5	Beyond SAIL: Enhancing SAIL via VNS and Parallel Computing	81
4.5.1	The V-SAIL Algorithm	83
4.5.2	The PV-SAIL Algorithm	84

4.6	Experimental Results	85
4.6.1	Experimental Setup	85
4.6.2	The Impact of Zero-Value Dilemma	88
4.6.3	The Comparison of SAIL and the Smoothing Technique	89
4.6.4	The Comparison of SAIL and Spherical K-means	91
4.6.5	Inside SAIL	91
4.6.6	The Performance of V-SAIL and PV-SAIL	94
4.7	Related Work	96
4.8	Concluding Remarks	96
	References	97
5	Selecting External Validation Measures for K-means Clustering	99
5.1	Introduction	99
5.2	External Validation Measures	100
5.3	Defective Validation Measures	101
5.3.1	The Simulation Setup	103
5.3.2	The Cluster Validation Results	104
5.3.3	Exploring the Defective Measures	104
5.3.4	Improving the Defective Measures	105
5.4	Measure Normalization	106
5.4.1	Normalizing the Measures	106
5.4.2	The Effectiveness of DCV for Uniform Effect Detection	112
5.4.3	The Effect of Normalization	114
5.5	Measure Properties	116
5.5.1	The Consistency Between Measures	116
5.5.2	Properties of Measures	118
5.5.3	Discussions	121
5.6	Concluding Remarks	122
	References	123
6	K-means Based Local Decomposition for Rare Class Analysis	125
6.1	Introduction	125
6.2	Preliminaries and Problem Definition	127
6.2.1	Rare Class Analysis	127
6.2.2	Problem Definition	128
6.3	Local Clustering	129
6.3.1	The Local Clustering Scheme	129
6.3.2	Properties of Local Clustering for Classification	129
6.4	COG for Rare Class Analysis	130
6.4.1	COG and COG-OS	130
6.4.2	An Illustration of COG	132
6.4.3	Computational Complexity Issues	133

6.5	Experimental Results	135
6.5.1	Experimental Setup	135
6.5.2	COG and COG-OS on Imbalanced Data Sets	137
6.5.3	COG-OS Versus Re-Sampling Schemes.	141
6.5.4	COG-OS for Network Intrusion Detection	141
6.5.5	COG for Credit Card Fraud Detection	145
6.5.6	COG on Balanced Data Sets	146
6.5.7	Limitations of COG	150
6.6	Related Work	151
6.7	Concluding Remarks	152
	References	152
7	K-means Based Consensus Clustering.	155
7.1	Introduction	155
7.2	Problem Definition	156
7.2.1	Consensus Clustering.	156
7.2.2	K-means Based Consensus Clustering	157
7.2.3	Problem Definition	158
7.3	Utility Functions for K-means Based Consensus Clustering	158
7.3.1	The Distance Functions for K-means.	159
7.3.2	A Sufficient Condition for KCC Utility Functions	159
7.3.3	The Non-Unique Correspondence and the Forms of KCC Utility Functions.	162
7.3.4	Discussions	163
7.4	Handling Inconsistent Data	165
7.5	Experimental Results	167
7.5.1	Experimental Setup	167
7.5.2	The Convergence of KCC	168
7.5.3	The Cluster Validity of KCC	169
7.5.4	The Comparison of the Generation Strategies of Basic Clusterings	171
7.5.5	The Effectiveness of KCC in Handling Inconsistent Data	173
7.6	Related Work	173
7.7	Concluding Remarks	174
	References	175
	Glossary	177