Prototype Based Clustering in High-Dimensional Feature Spaces

DISSERTATION

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.–Ing.)

angenommen durch die Fakultät für Informatik

der Otto–von–Guericke–Universität Magdeburg

von Dipl. Inform. Roland Winkler

geb. am 19.07.1982 in Bernburg

Gutachter

Prof. Dr. habil. Rudolf Kruse

Prof. Dr. habil. Frank Klawonn

P.D. Dr. habil. Alexander Hinneburg

Promotionskolloquium:

Magdeburg, den 18.06.2015

CONTENTS

1 INTRODUCTION 1
1.1 Examples . . . 1
1.2 Data Mining . . . 7
1.3 Motivation, Thesis Questions and Outline . . . 11

2 THE CURSE OF DIMENSIONALITY 15
2.1 Distance Concentration . . . 16
2.2 Conditions for Distance Concentration . . . 19
2.3 Distance Functions . . . 26
2.4 Tests for Distance Concentration . . . 28
2.5 Distribution Testing . . . 31
2.6 Noise, Overlapping Classes and Outliers . . . 32
2.7 Clustering in High-Dimensional Spaces . . . 37

3 PROTOTYPE BASED CLUSTERING ALGORITHMS 41
3.1 Mathematical Framework . . . 42
3.2 Hard c-Means . . . 48
3.3 Fuzzy c-Means . . . 50
3.4 FCM with Polynomial Fuzzifier Function . . . 54
3.5 Rewarding Crisp Memberships FCM . . . 60
3.6 Expectation Maximization . . . 65
3.7 Applying Prototype based Clustering Algorithms . . . 73

4 BENCHMARKS 83
4.1 Artificially Generated Data Sets . . . 84
4.2 External Cluster Quality Measures . . . 89
4.3 Internal Cluster Quality Measures . . . 93
4.4 Benchmark Setup . . . 98

5 EXPERIMENTAL RESULTS 103
5.1 Number of Dimensions and Clusters . . . 103
5.2 Comparing Clustering Algorithms . . . 119
5.3 Internal Cluster Quality Index Verification . . . 123

6 CONCLUSIONS AND FUTURE RESEARCH 137
6.1 Thesis Questions: Discussion . . . 137
6.2 Critique and Future Research . . . 141
6.3 Conclusions . . . 146

A ARTIFICIALLY GENERATED DATA SETS 151
A.1 Spherical Normal Shaped Classes of Identical Size . . . 151
A.2 Spherical Normal Shaped Classes of Various Sizes . . . 152

A.3 Distorted Classes Data Set 152
A.4 Corner Classes Data Set 163

B EDMOAL 167
B.1 Motivation for EDMOAL 167
B.2 The Basic Structure 169

C DATA AND RESULTS 171
C.1 Data Repository 171
C.2 Data Files 171
C.3 Result Files 172
C.4 Score List 174

D QUALITY OF CLUSTERING ALGORITHMS 177
D.1 Data Set Family D_1 178
D.2 Data Set Family D_2 180
D.3 Data Set Family D_3 182
D.4 Data Set Family D_4 184

E RANKING CAPABILITY OF INTERNAL INDICES 187
E.1 Data Set Family D_1 188
E.2 Data Set Family D_2 189
E.3 Data Set Family D_3 190
E.4 Data Set Family D_4 191

F INTERNAL AND THE F_1 INDEX CORRELATIONS 193
F.1 Data Set Family D_1 194
F.2 Data Set Family D_2 195
F.3 Data Set Family D_3 196
F.4 Data Set Family D_4 197

G INTERNAL VS. THE F_1 INDEX PLOTS 199
G.1 Plots for $m = 3$ and $c = 5$ 200
G.2 Plots for $m = 10$ and $c = 100$ 204
G.3 Plots for $m = 15$ and $c = 20$ 208
G.4 Plots for $m = 50$ and $c = 150$ 212

BIBLIOGRAPHY 217