
Contents

Motivation and objectives	XXXI
0.1 Motivation	XXXI
0.2 Objectives	XXXII
0.3 Main Contributions	XXXII
0.4 Publications	XXXIV
0.5 Document structure	XXXV

Part I The Smart Grid. Current scenario and future trends

1 Introduction	5
1.1 Liberalization of the energy market in the European Union	7
1.2 Liberalization of the energy market in Spain	9
1.3 Conclusions	11
2 The current power grid	13
2.1 The different roles involved in the management of Power Systems	13
2.2 The electrical energy network and the Smart Grid	20
2.3 Distributed Energy Resources (DER)	23
2.4 Renewable energies and their integration in the grid	29
2.5 Description of ancillary services	31
2.6 Conclusions	35
3 The impact of Advanced Metering Infrastructure and the Smart Grids	37
3.1 Standardisation in the European Smart Grids	37
3.1.1 The Smart Grid Architecture Model (SGAM)	39
3.1.2 The IEC-61850 standard	42
3.1.3 The Common Information Model (CIM)	43
3.2 The use of Big Data and Big Data Analytics in Power Systems	44

3.3	Demand Side Management and Demand Response	46
3.4	Conclusions and identification of needs	47
4	Conclusions and future trends	49

Part II State of the art on data mining and clustering techniques

5	An introduction to data mining	55
5.1	Data Mining and Knowledge Discovery in Databases	55
5.2	Data mining objectives	57
5.2.1	Descriptive data summarization	60
5.2.2	Class description	60
5.2.3	Frequent pattern mining	61
5.2.4	Classification	61
5.2.5	Cluster analysis	62
5.2.6	Outlier analysis	63
5.2.7	Correlation	64
5.2.8	Prediction	65
5.3	Techniques and algorithms for mining data	66
5.3.1	Fuzzy sets	66
5.3.2	Neural networks	71
5.3.3	Rough sets	76
5.3.4	Genetic and evolutionary algorithms	77
5.3.5	Case-based reasoning	79
5.3.6	Decision trees	79
5.3.7	Other methods	79
5.4	Introduction to dynamic data mining	80
5.4.1	Data stream mining	81
5.4.2	Time series data mining	84
5.4.3	Sequence data pattern mining	86
5.5	Conclusions	86
6	Description of clustering techniques	89
6.1	Introduction	89
6.2	Definition and types of data objects	90
6.3	Cluster distance measures	91
6.3.1	The Minkowski metric	91
6.3.2	Mahalanobis distance	93
6.3.3	Sample correlation coefficient	93
6.3.4	Matching coefficients	93
6.3.5	Entropy	94
6.3.6	Kullback-Leibler distance	95
6.3.7	Gowda and Diday distance	95
6.4	Normalization	96

6.5	Classification of clustering algorithms	96
6.6	Partitional clustering algorithms.....	98
6.6.1	Chain-map clustering	98
6.6.2	Max-min	99
6.6.3	K-means	100
6.6.4	PAM	100
6.6.5	CLARA	101
6.7	Hierarchical clustering algorithms.....	101
6.7.1	BIRCH	101
6.7.2	CURE	101
6.8	Fuzzy clustering algorithms	101
6.8.1	Definition of fuzzy partition	101
6.8.2	Fuzzy C-means (FCM)	103
6.8.3	Gustafson-Kessel or GK	104
6.8.4	Fuzzy Maximum Likelihood (FMLE).....	105
6.9	Density-based clustering algorithms	106
6.9.1	GDBSCAN	106
6.9.2	DENCLUE	106
6.10	Grid-based clustering algorithms	106
6.10.1	STING.....	106
6.10.2	CLIQUE	106
6.11	Spatial or geographical clustering algorithms	107
6.11.1	GRAVIClust	107
6.12	Clustering algorithms for distributed data	107
6.12.1	Collective Principal Component Analysis	107
6.12.2	RACHET	108
6.13	Quantitative and qualitative data clustering algorithms.....	108
6.13.1	K-modes and K-prototypes	108
6.13.2	ROCK	109
6.13.3	COOLCAT	109
6.13.4	Mixed-type variable fuzzy c-means (MVFCM)	110
6.14	Visualization and dimensionality reduction	110
6.15	Cluster validity indices	111
6.16	Conclusions	112
7	State of the art on clustering algorithms for time series	
data	115
7.1	Introduction	115
7.2	Distance or similarity measures for time series data	115
7.2.1	Dynamic Time Warping (DTW).....	116
7.2.2	Hausdorff distance.....	117
7.2.3	Edit Distance on Real sequences (EDR)	118
7.2.4	Time Warp Edit Distance (TWED).....	118
7.3	Classification of clustering algorithms for time series data.....	119

7.3.1	Classification according to the dynamic nature of data and the clustering algorithms	119
7.3.2	Classification according to the way the time series data is processed	121
7.3.3	Classification according to the number of features or characteristics of the data	121
7.4	State of the art on clustering algorithms for time series data ..	121
7.4.1	Raw time series data clustering	121
7.4.2	Feature-based time series data clustering	125
7.4.3	Model-based time series data clustering	128
7.5	Conclusions	130
8	Conclusions and proposal of development	135

**Part III Development of dynamic clustering techniques applied to
load profiles time series**

9	Introduction	145
9.1	Energy measures	145
9.2	Definition of load profile.....	145
9.3	Data mining objectives on load profiles	146
9.4	Conclusions	150
10	Assessment of data mining techniques for the analysis of load profiles	151
10.1	Techniques and algorithms for clustering and classification of load profiles	151
10.1.1	Cluster analysis and pattern recognition.....	152
10.1.2	Classification	154
10.2	Techniques and algorithms for forecasting of load profiles	157
10.2.1	Mathematical models based on quantitative and qualitative variables	158
10.2.2	Autoregressive models	160
10.2.3	Artificial Neural Networks (ANN) models	161
10.2.4	Self Organizing maps (SOM) models	162
10.2.5	Fuzzy inference models and expert systems	163
10.2.6	Statistical and probabilistic models	163
10.2.7	Hybrid models	164
10.3	Conclusions. Extending analysis to dynamic clustering	164
11	Development of algorithms and techniques to perform dynamic clustering on load profiles time series	167
11.1	Introduction	167
11.2	Data Model.....	169

11.3	Development of a data warehousing procedure for the pre-processing of time series data	171
11.4	Development of cluster validity indices for the evaluation of dynamic clustering on time series n-dimensional data	174
11.4.1	DB index	174
11.4.2	SD index	176
11.4.3	PC index	177
11.4.4	XB index	178
11.4.5	FS index	178
11.4.6	Summary of the dynamic distances needed	179
11.5	Development of a common framework for the dynamic clustering and visualization of daily load profile time series ...	179
11.5.1	Approach 1: comparing dimensions by the same time instant or day sample	180
11.5.2	Approach 2: evaluating dimensions as dynamic features	180
11.6	Development of a two-step time series clustering algorithm with a Hausdorff-based similarity distance for the dynamic clustering of daily load profile time series	184
11.6.1	Description of the two-step time series clustering algorithm	185
11.6.2	Decomposition of shapes in smaller linear surfaces	186
11.6.3	Similarity measure between surfaces based on the Hausdorff distance	187
11.6.4	Pseudocode	187
11.7	Conclusions	187
12	Application of the dynamic clustering framework to load profile time series from residential customers and results ..	189
12.1	Introduction. The GAD project	189
12.2	Database description	190
12.3	Sample of residential users used	191
12.4	Selection of the number of clusters to be found	193
12.5	Application of the developments made	193
12.5.1	First test: granularity options and dynamic k-means clustering. Analysis applied and results	194
12.5.2	Second test: common framework for dynamic clustering. Analysis applied and results	205
12.5.3	Third test: first approach of dynamic k-means clustering, common framework for dynamic clustering, and Hausdorff-based similarity measure dynamic clustering algorithm. Comparison and results	221
12.6	Conclusions from the tests	230

13 Conclusions from developments and tests	241
13.1 Data mining field of knowledge	241
13.2 Power systems field of knowledge	242

Part IV Conclusions

14 Main conclusions	247
15 Future Works	251
15.1 Type 4 dynamic clustering and batch processing of the time series data	251
15.2 Possibilistic clustering	252
15.3 Feature weighting or discrimination	252
15.4 Decomposition of the load profiles data objects in a variable number of smaller surfaces	253
15.5 Definition of indices for the trend of clusters	253
15.6 Dynamic clustering of quarter-hour energy demand measures ..	253
15.7 Application of dynamic clustering to the development of prediction models	254

Part V Appendices

A Work, energy and power	257
A.1 Work	257
A.2 Power	258
A.3 Energy	258
B Minimization of objective functions of clustering algorithms	259
B.1 Minimization of the K-means objective function	259
B.2 Minimization of the FCM objective function	260
C Linear Least Squares	263
D Cluster Validity Indices for Static Clustering	267
D.1 Partitional non-fuzzy clustering	267
D.1.1 Davies-Bouldin index	267
D.1.2 Dunn and Dunn-like indices	268
D.1.3 RMSSDT and RS indices	268
D.1.4 SD validity index	269
D.2 Partitional fuzzy clustering	270
D.2.1 PC index	270
D.2.2 PE index	270
D.2.3 XB index	271
D.2.4 Fukuyama - Sugeno index	271

Contents XXIII

D.2.5 Gath and Geva indices	272
D.3 Collection of indices	272
E Methods to Determine Sample Size	275
E.1 Combination of samples from a given population	275
E.2 Definition of confidence margin	275
E.3 Population parameters	276
E.4 Sample parameters	276
E.5 Sample size with sampling error and confidence margin	277
E.6 Stratified sample	278
E.7 Example of sample sizes: database of electric energy consumption	278
References	281

List of Figures

1.1 Example of market clearing price in the Spanish Energy Market Pool. Source: OMIE (Operador del Mercado Ibérico de Energía).	7
2.1 Diagram of the Spanish power distribution, with the typical voltage levels in generation, transmission and distribution. Source: Red Eléctrica de España (REE), the Spanish TSO. Translated to English by Ignacio Benítez.	21
2.2 Smart Grid applications in the future schema of electricity generation, transmission and distribution. Source: MyISM3004 blog and Sean Dempsey, 2011.	23
2.3 Technology areas in the Smart Grids. Source: Smart Grids Roadmap, International Energy Agency, 2011.	24
3.1 The SGAM (Smart Grid Architecture Model). Source: Heise.de.	40
5.1 Classification of Data mining techniques according to the knowledge mined.	59
5.2 Example of two clusters (A and B) and their respective centroids, (x_1, y_1) and (x_2, y_2) , on a 2D data set.	63
5.3 Example of fuzzification process: three membership functions are defined for the variable “weight”.	67
5.4 Example of application of hedges “very” and “more or less” on membership function “hot”.	68
5.5 Example of Mamdani inference: three membership functions are defined for the output variable “BMI”.	68
5.6 Example of Mamdani inference: description of inference rules.	69
5.7 Example of Mamdani inference: defuzzification process.	70
5.8 Example of Sugeno inference.	70
5.9 Mathematical representation of a biological neuron and synapse process.	72

XXVI List of Figures

5.10 Sigmoid function.....	73
5.11 Example of prediction from multilayer ANN.....	74
5.12 Multilayer ANN.	74
5.13 Spatial assignment of energy consumption load profiles in SOM neurons based on similarity.	76
5.14 Resulting prototypes after applying the K-means clustering algorithm on the SOM neuron codebooks.	77
5.15 CIPT (Charging Infrastructure Planning Tool). FP7 Project MOBINCITY (Smart Mobility in Smart City), grant agreement No. 314328.	78
6.1 Minkowski metrics.	92
6.2 Classification of clustering algorithms.	113
10.1 Clustering algorithm applied on the trained SOM.	156
10.2 Prototypes obtained from the SOM classification.	157
11.1 Data cell structure after preprocessing.	170
11.2 Matrix of centroids (c, c_{ini}).	171
11.3 Membership matrix (u).	172
11.4 Exclusive membership matrix ($pertain$).	172
11.5 Data preprocessing stage.	173
11.6 Computation of distances between objects and clusters.	181
11.7 Computation of membership matrix u	185
12.1 DB indices for daily load profiles form year 2008	194
12.2 Cluster prototypes from dynamic clustering on sequence of daily load profiles.	197
12.3 Cluster prototypes from dynamic clustering on sequence of daily load profiles, working days.	199
12.4 Cluster prototypes from dynamic clustering on sequence of daily load profiles, non-working days.	201
12.5 Cluster prototypes from dynamic clustering on sequence of cumulated monthly daily load profiles.	203
12.6 The six classes in the UCI Synthetic Control Chart Series.	207
12.7 Rearrangement of the UCI Synthetic Control Chart Series as a dataset with six classes of 10 objects each, with 60 time instants.	208
12.8 FFCM cluster prototypes on the UCI Synthetic Control Chart Series.	209
12.9 END-KME clustering. Cluster prototypes on the UCI Synthetic Control Chart Series.	210
12.10END-KMC clustering. Cluster prototypes on the UCI Synthetic Control Chart Series.	211
12.11END-FCMC. Cluster prototypes on the UCI Synthetic Control Chart Series.	212

List of Figures XXVII

12.12END-FCME clustering. Cluster prototypes on the UCI Synthetic Control Chart Series.	213
12.13FFCM cluster prototypes on the dataset of daily load profiles.	216
12.14END-FCMC clustering. Cluster prototypes on the dataset of daily load profiles.	217
12.15END-KMC clustering. Cluster prototypes on the dataset of daily load profiles.	218
12.16END-KME clustering. Cluster prototypes on the dataset of daily load profiles.	219
12.17END-FCME clustering. Cluster prototypes on the dataset of daily load profiles.	220
12.18Raw data being analyzed.	222
12.19Results from END-KME dynamic clustering algorithm.	232
12.20Results from END-KMC dynamic clustering algorithm.	233
12.21Results from END-KMH dynamic clustering algorithm.	234
12.22Results from Extended K-means clustering algorithm.	235
12.23Results from END-FCME dynamic clustering algorithm.	236
12.24Results from END-FCMC dynamic clustering algorithm.	237
12.25Results from END-FCMH dynamic clustering algorithm.	238
12.26Results from FFCM dynamic clustering algorithm.	239
C.1 Example of Least Squares applied on a dynamic Load Curve.	264
C.2 Example of Least Squares applied on a dynamic Load Curve. Histogram of residuals.	265

List of Tables

6.1	Clustering algorithms by main characteristic and similarity measure used.	114
7.1	Types of clustering according to dynamic nature of data and classes	120
7.2	Summary of raw data dynamic clustering algorithms found in the literature.	131
7.3	Summary of feature-based dynamic clustering algorithms found in the literature.	132
7.4	Summary of model-based dynamic clustering algorithms found in the literature.	133
11.1	Formation of different Type 3 dynamic clustering algorithms based on Weber description.	183
12.1	Objects data set. Description of columns or variables.	191
12.2	Clusters obtained from dynamic clustering on sequence of daily load profiles.	196
12.3	Clusters obtained from dynamic clustering on sequence of daily load profiles, working days.	200
12.4	Clusters obtained from dynamic clustering on sequence of daily load profiles, non-working days.	200
12.5	Clusters obtained from dynamic clustering on sequence of cumulated monthly daily load profiles.	202
12.6	Summary of results of dynamic clustering techniques applied on the UCI Synthetic Control Chart Series.	214
12.7	Summary of results of dynamic clustering techniques applied on the dataset of daily load profiles from residential customers.	215
12.8	Type 3 dynamic clustering algorithms tested.	223
12.9	Test results, DB modified index.....	224
12.10	Test results, SD modified index.	224

XXX List of Tables

12.11	Test results, PC index.	225
12.12	Test results, XB modified index.	225
12.13	Test results, FS modified index.	225
12.14	Assignment of clusters from END-KME to expected groups.	228
12.15	Assignment of clusters from END-KMC to expected groups.	228
12.16	Assignment of clusters from END-KMH to expected groups.	229
12.17	Assignment of clusters from Extended K-means to expected groups.	229
12.18	Assignment of clusters from END-FCME to expected groups.	229
12.19	Assignment of clusters from END-FCMC to expected groups.	229
12.20	Assignment of clusters from END-FCMH to expected groups.	229
12.21	Assignment of clusters from FFCM to expected groups.	230
13.1	Assignment of clusters to expected groups, results from Test 3 . . .	242
13.2	Assignment of clients to expected groups, results from Test 3 . . .	242
D.1	Cluster validity indices, for fuzzy and non-fuzzy partitions.	273