

Intelligent Analysis of Small Data Sets for Food Design

David Peter Alfred Corney

Department of Computer Science
University College, London

PhD
September 2002

Abstract

This thesis compares the performance of machine learning techniques and statistics in the analysis of food design data. The goal of the analysis is to understand what makes people like (or dislike) a product, by building models relating sensory features (such as flavour or texture) to consumer preferences. One difficulty in analysing these data sets is that they are extremely small, due to “taste-fatigue” of consumer preference panels.

Feature selection is essential because food sensory data sets typically have many features and few records. Several feature selection algorithms are compared, and the results highlight the need to limit the number of features used. We therefore apply model order selection to feature selection. A semi-supervised feature selection method is introduced and compared with more traditional methods.

After the selection of a suitable set of features, the relationship between those features and consumers’ preferences must be modelled. Two regression techniques are compared, focussing on their relative performance on very small data sets. A semi-supervised ensemble learning algorithm is introduced, and analysed.

Consumers have individual preferences, so rather than producing a single generic product, food designers must first discover homogeneous groups of consumers, and then target each group with a different product. Several clustering techniques are compared, and consideration of their inherent biases reveals further information regarding the structure of the data. A combination of regression and clustering is proposed, which allows evaluation of clustering results using the predictive power of the resultant models.

Preference data sets contain a significant number of misleading outliers owing to the way they are collected. An algorithm that combines clustering and outlier detection is introduced, which aims to produce an outlier-free cluster model, and also provides heuristic estimates of the number of outliers present.

Overall, machine learning techniques show performance similar to traditional statistical techniques, with small improvements in accuracy in some cases. Machine learning brings the benefit of typically being dependent on fewer assumptions: where these assumptions are invalid, results may be improved. Furthermore, machine learning makes use of considerable computational power, which is now cheaply available, in the search for improved solutions. In this thesis, we examine the efficacy of machine learning techniques when analysing food design data sets.

In summary, the main contributions of this thesis are:

- A semi-supervised feature selection algorithm
- A semi-supervised ensemble for regression
- A clustering evaluation technique
- An outlier detection technique for clustering

Acknowledgements

I am fortunate to have been surrounded by many interesting and helpful people at UCL, Sira and Unilever. Special thanks are due to Rob Burbidge and Matt Trotter for many beneficial discussions, often under John Smith's influence. My work would have been more difficult and much less fun without all three of them.

The Primal and Nuclear research groups at UCL have been a frequent source of inspiration, not least due to Bill Langdon, Peter Bentley, Hugh Mallinson and Sean Holden.

My supervisors, John Gilby (Sira Ltd), Philip Treleaven (UCL) and Bernard Buxton (UCL), have consistently provided good advice from day one, without which I would have struggled, and would still be struggling. I would also like to thank Steve Wooding, Keith Plater and Mark Singleton, and everyone in the Data Sciences group of Unilever Research, for their invaluable guidance throughout this work, and for the provision of the data sets, small as they were.

The research was undertaken within the Postgraduate Training Partnership established between Sira Ltd and University College, London. Postgraduate Training Partnerships are a joint initiative of the Department of Trade and Industry and the Engineering and Physical Sciences Research Council. The work has been generously sponsored by Unilever Research plc.

— David Corney
October 2001

Contents

1	Intelligent Food Design	11
1.1	Why design food?	11
1.2	Intelligent systems, product design, and the food industry	11
1.2.1	Product design	12
1.2.2	Food and drink	14
1.2.3	Intelligent systems	15
1.2.4	Intelligent systems for product design	17
1.2.5	Designing food products	18
1.2.6	Intelligent systems in the food industry	21
1.2.7	Intelligent systems for food design	22
1.3	Data analysis	23
1.4	Data gathering	24
1.4.1	Sensory panels	24
1.4.2	Preference panels	25
1.4.3	Other issues	26
1.4.4	Description of data sets	28
1.5	Thesis outline	29
1.5.1	Feature selection and regression	30
1.5.2	Cluster analysis	31
1.5.3	Outlier detection	32
1.6	Conclusions	32
2	Feature Selection and Regression	34
2.1	Introduction	34
2.1.1	Selecting sensory features	35
2.1.2	Relevant features	37
2.2	Regression	37
2.2.1	Linear regression	38
2.2.2	Radial basis function networks	39
2.3	Approaches to feature selection	40
2.4	Searching with wrappers	43
2.4.1	Exhaustive search	43
2.4.2	Sequential methods	44
2.4.3	Simulated annealing	46
2.4.4	Hybrid methods	46
2.4.5	Degree of search exhaustion	47

2.5	Avoiding overfitting	48
2.5.1	Hold out methods	51
2.5.2	Complexity penalisation	54
2.5.3	Model combination	56
2.6	Evaluation of feature sets	58
2.7	Results on food data sets	59
2.8	Semi-supervised feature selection	62
2.8.1	Method	63
2.8.2	Results on meat data	63
2.9	Minimum description length for regression	66
2.10	The EM algorithm for regression	67
2.11	Semi-supervised ensembles	68
2.11.1	Semi-supervised bagging for Boston Housing data	70
2.11.2	RBF network size heuristic	72
2.11.3	Correlations and distances	73
2.11.4	Committee sizes	74
2.11.5	Committee effectiveness	75
2.11.6	Varying the amount of unlabelled data	77
2.11.7	Food data results	78
2.11.8	SSEL results summary	79
2.12	The effect of the initial number of features	80
2.13	Conclusions	82
3	Cluster Analysis	85
3.1	Introduction	85
3.2	Clustering preference data	86
3.3	Similarity and distance measures	87
3.4	Clustering methods	88
3.4.1	Preference mapping	88
3.4.2	Locally linear preference mapping	91
3.4.3	Preference mapping summary	92
3.4.4	k -means clustering	93
3.4.5	Gaussian mixture models and the EM algorithm	94
3.4.6	Biases in clustering algorithms	97
3.5	Comparing and validating segmentations	98
3.5.1	The Rand index	99
3.5.2	External criterion	101
3.5.3	Train and test	101
3.5.4	Replication analysis	102
3.5.5	Hypothesis testing	103
3.5.6	Robustness	103
3.5.7	Summary of cluster validation methods	104
3.6	How many clusters?	104
3.6.1	Existence of clusters	105
3.6.2	Information criteria	106
3.6.3	The effect of small data sets	109
3.6.4	Cluster shapes	110

3.7	Which algorithm?	112
3.8	Consistency versus accuracy	113
3.9	Usefulness of predictions	117
3.9.1	Weighted cluster-regression	119
3.10	Conclusions	121
4	Outlier Detection	123
4.1	Introduction	123
4.1.1	Outliers in food preference data	124
4.1.2	Actions on outliers	125
4.2	Influential points	125
4.3	Measuring the accuracy of outlier detectors	126
4.4	Accommodation and rejection	127
4.4.1	Accommodation	128
4.4.2	Discordancy	129
4.4.3	Summary: accommodation vs. discordancy	133
4.5	Clustering and outliers	133
4.5.1	Finding outliers in preference data	135
4.5.2	Discordancy experiments	135
4.6	Sequential outlier rejection in clustering	137
4.6.1	The SOR algorithm	138
4.6.2	Sequence consistency	140
4.6.3	How many outliers?	141
4.6.4	Convex hull inclusion test	144
4.6.5	Validation methods	144
4.6.6	Outlier detection results	145
4.6.7	SOR for food data	147
4.6.8	SOR with weighted cluster-regression	148
4.7	Conclusions	149
5	Future work	153
5.1	Introduction	153
5.2	Alternative routes to intelligent food design	153
5.2.1	Virtual consumers	153
5.2.2	Consumer committees	154
5.2.3	Grouping consumers drivers	155
5.2.4	Grouping product drivers	155
5.2.5	Virtual products	155
5.2.6	Separate products	156
5.2.7	Separate products and virtual consumers	156
5.3	Alternative methods for data analysis	157
5.3.1	Semi-supervised feature selection	157
5.3.2	Extending cross model validation	157
5.3.3	Fuzzy Rand index	157
5.3.4	Analysis of GMM components	159
5.3.5	Catalogue design	159
5.3.6	Sequential outlier rejection for classification and regression	160

5.3.7	Sequential outlier rejection for feature selection	161
5.3.8	Replacement machine learning methods	161
5.4	Alternative applications	162
5.4.1	Mass customisation of consumer goods	163
5.4.2	Collaborative filtering for food	163
5.5	Data gathering	164
5.6	Conclusions	164
6	Conclusions	165
6.1	Chapter summary	165
6.2	Conclusions	166
6.3	Contributions	167
A	Experimental details	168
A.1	Feature selection and regression	168
A.1.1	Radial basis functions	168
A.1.2	Simulated annealing	168
A.1.3	Semi-supervised feature selection	169
A.2	Outliers in n -dimensional Gaussians	169
A.3	Further SSEL results	170

List of Figures

1.1	Three research domains	12
2.1	Filter methods for feature selection	41
2.2	Correlation based feature elimination	42
2.3	Wrapper methods for feature selection	42
2.4	Space of feature subsets	43
2.5	Two stages of overfitting with wrappers	49
2.6	Model order selection hierarchy	50
2.7	Bias when splitting data set	51
2.8	Cross model validation algorithm	53
2.9	Semi-supervised feature selection	63
2.10	Semi-supervised feature selection — linear regression error for meat data	65
2.11	Semi-supervised feature selection — linear model size for meat data	65
2.12	Semi-supervised feature selection — RBF error for meat data	65
2.13	Semi-supervised feature selection — RBF model size for meat data	65
2.14	MDL scores for linear regression on meat data	67
2.15	MDL scores for linear regression on meat data — error and complexity	67
2.16	EM regression algorithm	68
2.17	EM algorithm on RBF networks for meat data	68
2.18	EM algorithm on RBF networks for small Boston sample	68
2.19	Semi-supervised ensemble learning (SSEL) algorithm	69
2.20	SSEL error with 20 records, Boston data	75
2.21	SSEL error with 50 records, Boston data	75
2.22	SSEL error with 100 records, Boston data	75
2.23	SSEL error, meat data	75
2.24	Feature selection error, various feature set sizes	81
2.25	Number of features selected, various feature set sizes	81
3.1	Normalised preference map, artificial data	90
3.2	Scaled preference map, artificial data	90
3.3	Normalised preference map, meat preference data	90
3.4	Scaled preference map, meat preference data	90
3.5	Locally linear embedding algorithm	91
3.6	k -means clustering algorithm	93
3.7	General EM algorithm	95
3.8	Description length for meat preference data, spherical GMM	107
3.9	Complexity and likelihood for meat preference data	107

3.10	Description length for vegetable preference data, spherical GMM	108
3.11	Complexity and likelihood for vegetable preference data	108
3.12	Description length for beverage preference data, spherical GMM	108
3.13	Complexity and likelihood for beverage preference data	108
3.14	Description length for artificial data with various data set sizes	109
3.15	Description length for artificial data with non-spherical components	109
3.16	Pathological fitness landscape	114
3.17	Mutli-modal fitness landscape	114
3.18	Accuracy and consistency: artificial data generated from four Gaussian components.	115
3.19	Accuracy and consistency: vegetable preference data	115
3.20	Accuracy and consistency: beverage preference data	115
3.21	Accuracy and consistency: meat preference data	115
3.22	Clustering with regression: beverage data	120
3.23	Clustering with regression: meat data	120
3.24	Clustering with regression: vegetable data	120
4.1	Variance inflation: inlier and outlier distributions	130
4.2	Artificial regression data without outliers	131
4.3	Artificial regression data with one outlier	131
4.4	Two-cluster model	134
4.5	One-cluster model	134
4.6	Iterative Gaussian discordancy, one cluster	137
4.7	Iterative Gaussian discordancy, two clusters.	138
4.8	Sequential outlier rejection (SOR) algorithm	139
4.9	SOR sequence	139
4.10	SOR error rate	139
4.11	Error rate for unimodal Gaussian data with 80 added outliers	142
4.12	Error rate with second difference	142
4.13	Error rate with smoothed second derivative	143
4.14	Error rate for beverage preference data, with smoothed second derivative	143
4.15	SOR error for uniform data with no outliers	143
4.16	SOR error for Gaussian data with 50 outliers	143
4.17	SOR algorithm for weighted cluster-regression	149
4.18	Weighted cluster regression with SOR, meat data	150
4.19	Weighted cluster regression with SOR, vegetable data	150
4.20	Weighted cluster regression with SOR, beverage data	150
5.1	Modified cross model validation algorithm	158
5.2	Sequential Feature Rejection Algorithm	162

List of Tables

2.1	Feature selection results — meat data. Key in Table 2.3.	60
2.2	Feature selection results — vegetable data. Key in Table 2.3	60
2.3	Key to Tables 2.1-2.2	61
2.4	Semi-supervised ensemble results — small Boston sample	72
2.5	Semi-supervised ensemble results — medium Boston sample	72
2.6	Semi-supervised ensemble results — large Boston sample	72
2.7	Semi-supervised ensemble distances — Boston data	74
2.8	Semi-supervised ensemble correlations — Boston data	74
2.9	Committee effectiveness — Boston data	76
2.10	Boston data with varying labelled and unlabelled record set sizes	77
2.11	Semi-supervised ensemble results — meat data	79
3.1	Feature extraction results: LLE vs. PCA	92
3.2	Contingency table comparing two GMM partitions of the beverage preference data	99
3.3	Biases in Gaussian mixture models	111
3.4	Clustering algorithm comparison	113
3.5	Relative accuracy of weighted cluster regression	121
4.1	Correlations of repeated SOR sequences	141
4.2	Convex hull inclusion test results	144
4.3	Estimated number of outliers (low dimensional data)	146
4.4	Estimated number of outliers (high dimensional data)	147
4.5	Estimated number of outliers in preference data sets	148
A.1	Estimated Gaussian width factors	170
A.2	Auto-mpg data with varying labelled and unlabelled record set sizes . .	171
A.3	Abalone data with varying labelled and unlabelled record set sizes . . .	171

List of Abbreviations

RBF	Radial basis function	39
FSS	Forward sequential selection	44
ADJ	Adjusted distance metric	63
SSEL	Semi-supervised ensemble learning	68
PCA	Principal components analysis	88
LLE	Locally linear embedding	91
GMM	Gaussian mixture model	94
EM	Expectation maximisation	94
MDL	Minimum description length	106
SOR	Sequential outlier rejection	137