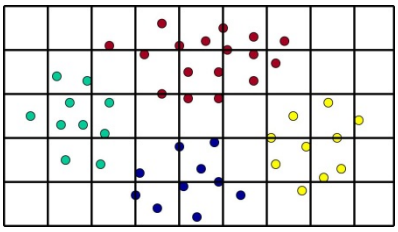
G-COM

Galor Cluster Optimization

The solution for data structure analysis

Data Mining and Prediction: The predominant task in the analysis of massive data is the identification of (statistically) relevant structures and a subsequent valid prediction of future developments.

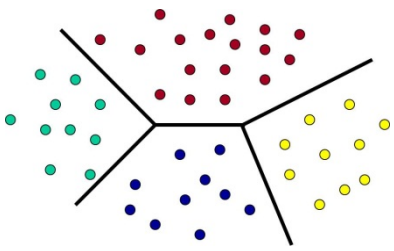


Sparsely populated and empty cells

Classical Methods: General classical approaches are often based on low dimensional pre-segmentation techniques for creating near homogenous substructures that are then subject to statistical analysis. In the presence of a reasonably large number of parameter characteristics and specifications the generated cells are typically sparsely populated. Hence, quite often, an application of the law of large numbers is prohibited and different, more complicated and less reliable, statistical techniques have to be evoked.

A Change of Paradigm: From a theoretical point of view there is an obvious alternative: Rather than using a simple dissection of parameter space followed by complicated statistics one can utilize an optimal clustering of this space followed by subsequent simple, meaningful and reliable statistics. Until recently an effective and efficient application of this natural principle was out of reach due to the lack of an adequate mathematical model and fast algorithms for clustering of high dimensional weighted data, including nominal data, under all relevant problem specific constraints. While standard clustering methods care capable of determining homogenous structures efficiently, such constraints could not been incorporated appropriately. And indeed, typically these constraints allow to gain new insight in the respective application by allowing for a significant analysis of the clusters. Further, standard methods are typically restricted to computing solutions that are merely locally optimal.

Galor Cluster Optimization Model (G-COM):



Significant substructures

In many years of research, Dr. Andreas Brieden und Dr. Peter Gritzmann succeeded in developing G-COM to transform this natural idea into a scientifically justified sound and practically highly efficient method. G-COM first solves an application and data specific complex clustering problem with the aid of state-of-the-art and partly newly developed mathematics and software technology. Based on the detected structures within the data a robust and statistically significant prediction method is then applied to the quite homogenous substructures.

Fields of Application: G-COM can be applied to a great variety of data analysis tasks in various business sectors. It has already been proven highly successful in the prediction of insurance loss, credit defaults and air cargo demands.

Prerequisites: G-COM detects hidden structures in massive data. Naturally, the available data must contain the parameters and variables that are relevant for the desired prediction.
For a proof of concept for the potential of G-COM in a new field of application the following is needed:

- A short description of the application including the desired analytic goals;
- A list and specifications of the relevant parameters;
- A representative set of test data.

Gesellschaft für Angewandte Logik und Operations Research • Spitzelbergstraße 10a, D-81476 Munich, Germany
Managing Directors: Dr. Peter Gritzmann, Dr. Andreas Brieden
Phone ++49 / 89 / 74 54 53 70 • Fax ++49 / 89 / 74545380 • www.galor.de • office@galor.de

Examples:

- **Prediction of damage-frequency in housing insurance**

  Target value: Probability of observing damage within the next year

  Explaining variables: Water Zone
  Year of constructing
  Size of building
  …

Cluster	1	2	3	4	5	...	All
Frequency	**0,004%**	**0,008%**	**0,010%**	**0,020%**	**0,030%**		**0,012%**
Water-Zone							
1 (best)	15,00%	100,00%	0,00%	0,00%	10,00%		15,00%
2	85,00%	0,00%	0,00%	70,00%	45,00%		40,00%
3	0,00%	0,00%	90,00%	20,00%	30,00%		30,00%
4 (worst)	0,00%	0,00%	10,00%	10,00%	10,00%		15,00%
Sum	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**100,00%**		**100,00%**
Year of Construction							
1 (newest)	0,00%	30,00%	0,00%	10,00%	15,00%		10,00%
2	0,00%	20,00%	0,00%	10,00%	10,00%		10,00%
3	0,00%	5,00%	0,00%	30,00%	5,00%		10,00%
4	0,00%	5,00%	0,00%	20,00%	5,00%		10,00%
5	0,00%	0,00%	0,00%	20,00%	10,00%		10,00%
6	0,00%	10,00%	0,00%	10,00%	10,00%		10,00%
7	10,00%	30,00%	20,00%	0,00%	15,00%		10,00%
8	25,00%	0,00%	20,00%	0,00%	5,00%		10,00%
9	25,00%	0,00%	30,00%	0,00%	5,00%		10,00%
10 (oldest)	40,00%	0,00%	30,00%	0,00%	20,00%		10,00%
Sum	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**100,00%**		**100,00%**
Size							
...							

How to read? E.g., buildings in Cluster 2 have an average damage frequency of 0,008%, all belong to the best water zone and 30% belong to the 10% newest buildings.

Lessons learned: E.g., increasing age does not mean increasing risk!
E.g., worse water zone does not mean increasing risk!
…
And actuarials can explain why; it depends on the combination of attributes!

Implication: E.g., current premiums can be corrected by a factor that simultaneously accounts for multivariate correlation.

Results: Current premium systems are outperformed in economic scenario analyses.

Gesellschaft für Angewandte Logik und Operations Research • Spitzelbergstraße 10a, D-81476 Munich, Germany
Managing Directors: Dr. Peter Gritzmann, Dr. Andreas Brieden
Phone ++49 / 89 / 74 54 53 70 • Fax ++49 / 89 / 74545380 • www.galor.de • office@galor.de

- **Prediction of efficacy of different medications**

  Target value: Probability to respond to applied medication

  Explaining variables: Medication
  Region
  Body Mass Index (BMI)
  Age
  Gender
  …

Cluster	1	2	3	4	5	…	All
Response Probability	**50%**	**20%**	**60%**	**40%**	**70%**		**50%**
Medication							
dose rate 10mg	60%	50%	65%	0%	0%		35%
dose rate 20mg	15%	30%	10%	0%	85%		19%
dose rate 25mg	25%	20%	25%	0%	15%		15%
placebo	0%	0%	0%	100%	0%		31%
Sum	**100%**	**100%**	**100%**	**100%**	**100%**		**100%**
Region							
eu	20%	20%	25%	20%	35%		25%
non eu	80%	80%	75%	80%	65%		75%
Sum	**100%**	**100%**	**100%**	**100%**	**100%**		**100%**
BMI							
1 (lowest)	25%	25%	20%	15%	20%		20%
2	20%	25%	20%	20%	20%		20%
3	20%	15%	20%	20%	25%		20%
4	15%	15%	20%	20%	20%		20%
5 (highest)	20%	20%	20%	25%	15%		20%
Sum	**100%**	**100%**	**100%**	**100%**	**100%**		**100%**
Sex							
male	30%	40%	35%	35%	30%		35%
female	70%	60%	65%	65%	70%		65%
Sum	**100%**	**100%**	**100%**	**100%**	**100%**		**100%**
Age							
…							

How to read? E.g., 60% of patients in Cluster 3 have responded to medication, 65% of them have been treated with dose rate 10mg.

Lessons learned: Response on patients does not depend on single items, it depends on the combination of attributes!
And medical scientists can explain why.

Implication: For any patient the response probability can be predicted depending on her or his attribute combination.

Results: Predictions (based on statistical testing) for efficacy of medication has been proved to be highly significant.

Further information/contact: g-com@galor.de