Applying “Emergent Law based Statistics” to prediction and decision tasks

by André Kuck, Hans Frischhut and Norbert Kratz

As a starting point in the development of emergent law based statistics and the derived decision theory we used the following fundamental principles:

- Empirical knowledge is defined as patterns of measurements that were always observed.
- To measure the predictive features of empirical knowledge of time t it is evaluated at time $t + T$. So it was possible to find general meta-patterns that show the stability of knowledge in the course of time.
- The problem of decision theory is seen as the search for sequences of more elementary behavior that always lead to superior overall results – instead of trying to find optimal single actions.

It came as a surprise how far reaching the consequences of obeying these principles are. We found superior performance in predictive modelling and algorithmic decision making compared to stochastic approaches. And we found applications that are simply not possible in the world of stochastic modelling. An important example is the transfer of knowledge between databases. Therefore, we are now in the process of extracting knowledge from public sources to generate “Commonsense Knowledge” about always observed properties and processes of the world.

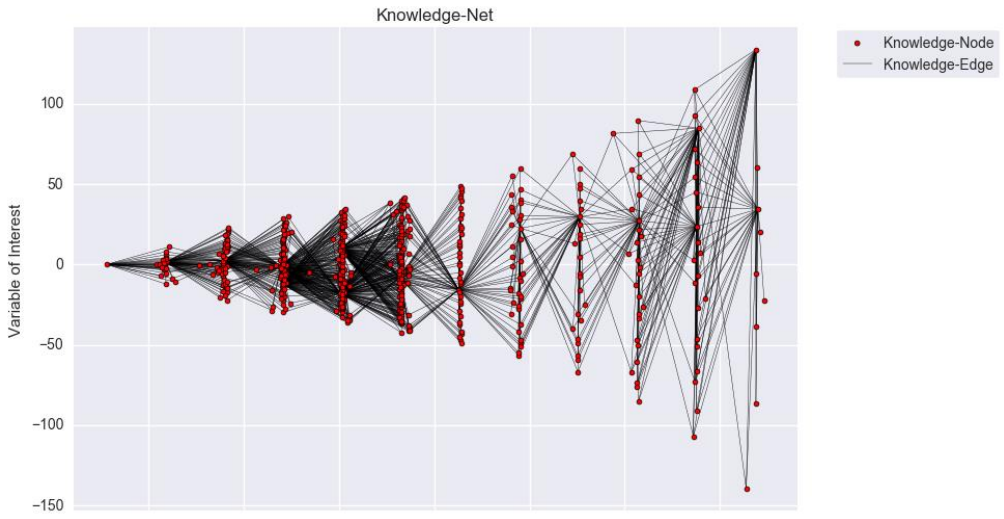
As our approach results from a fundamental change in statistical methodology and gives access to “Commonsense Knowledge”, it is applicable to all areas where predictions are needed to choose the best behavioral strategies:

- Marketing, Pricing, Risk Management, Asset Management etc.
- Logistics, Management of technical Processes, Quality Management etc.

Solving classical prediction problems by “KnowledgeNet” based models

The first step in the emergent law based model building process extracts “**KnowledgeNets**” from databases.

Because we define knowledge empirically as “until now always repeated patterns” we are able to create directed graphs that show all relations of objects concerning one or more variables of interest. Our approach enables us to ensure that all interesting emergent laws of a certain quality hidden in the data can be found. In this sense we can make sure to find **”complete knowledge”**.



Graph 1: The Structure of a KnowledgeNet created for the variable of interest “Number of rented bicycles in the next hour”

Video of the development of the KnowledgeNet: http://www.udpl.de/knowledgenet.html

The nodes in the KnowledgeNet show objects (groups of observations) that always had a certain relation – a greater (or smaller) number of rented bicycles – to all other nodes that are connected by edges. So all the edges in the KnowledgeNet represent emergent laws.

KnowledgeNets are our empirical basis for the construction of predictive models.

- The resulting models reach a prediction accuracy that is at least comparable with up to date machine learning algorithms.
- Feature creation and selection is performed automatically by the learning process.
- On the basis of a single KnowledgeNet models for several variables of interest can be constructed.
- This also works for variables that were not used to create the net.
- The resulting models only consist of emergent laws – every relation used to predict the variable of interest was always true. No assumptions are needed.
- The models are no black boxes and the used laws can be perfectly understood by the user.
- The approach can handle the falsification of laws because a KnowledgeNet contains enough objects to replace falsified laws by laws attached to other objects.
- The approach to use only emergent laws is empirically justified by meta-laws. (see http://udpl.eu/files/HansFrischhut_BA_Internet.html)

Prediction task	Best Mean Absolute Prediction Error of some state-of-the-art Machine-Learning algorithms	Mean Absolute Prediction Error of Emergent Law based Model
Number of rented bicycles	33.77	27.9
Temperature in Washington DC	0.01343	0.01278
Humidity in Washington DC	0.03643	0.0359

Table 1: prediction performance for several variables of interest created from the above KnowledgeNet.

Improving decision results by using KnowledgeNets, T-Dominant decision rules and "World-Views"

To make predictions valuable they need to be connected to a decision problem. In our approach the concept of **T-Dominance** lies at the heart of decision making. T-Dominance simply states:

- If you have the choice between two decision rules, select the one that led to better results after each sequence of T decisions.

For example, if we want to invest money in loans arranged by the platform "Lending Club" we could evaluate different loan selection strategies after every sequence of 10,000 loans empirically:

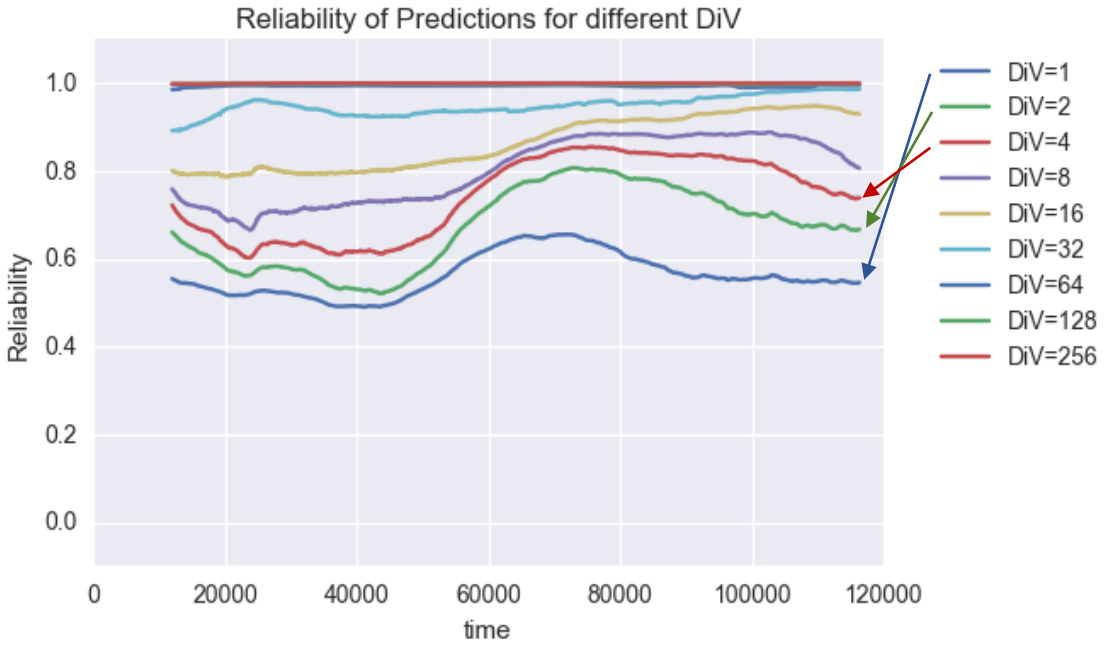
Strategy	$r_{t,min}$(T=10,000)	$r_{t,max}$(T=10,000)	SR	RAROC(0.001)	r_{t+T} – Evaluation
Only select A-Rated	6.99%	11.23%	15.8	5.43	8.22%
Only select C-Rated	12.36%	16.86%	13.8	2.49	14.03%

Table 2: Evaluation of different loan selection strategies.

The common stochastic decision criteria (Sharp Ratio (SR) and Risk adjusted Return on Capital (RAROC)) lead to the choice of only selecting A-rated loans.
But in each sequence of 10,000 loans the return of C-rated loans was superior up to time *t* (columns 2 and 3 in the above table) – the strategy "only select C-rated loans" T-dominated the strategy "only select A-rated loans". The last evaluation column shows that in the following 10,000 loans the advice

of T-Dominance leads to better results (about 6% higher return) compared to stochastic decision criteria.

The general use of T-Dominant decision rules is justified by emergent meta-laws about the predictability of the performance of different selection strategies:



Graph 2: Development of the rate of true predictions (Reliability) for predictions about the relative performance of loan selection strategies with different DiV. (Bottom blue line is $DiV = 1$, the green line above is $DiV = 2$, the red line above is $DiV = 4$)

The above graph shows the rate of true predictions about the relative performance of loan portfolios made with different selection strategies with different DiV (the number of verifications is called DiV).

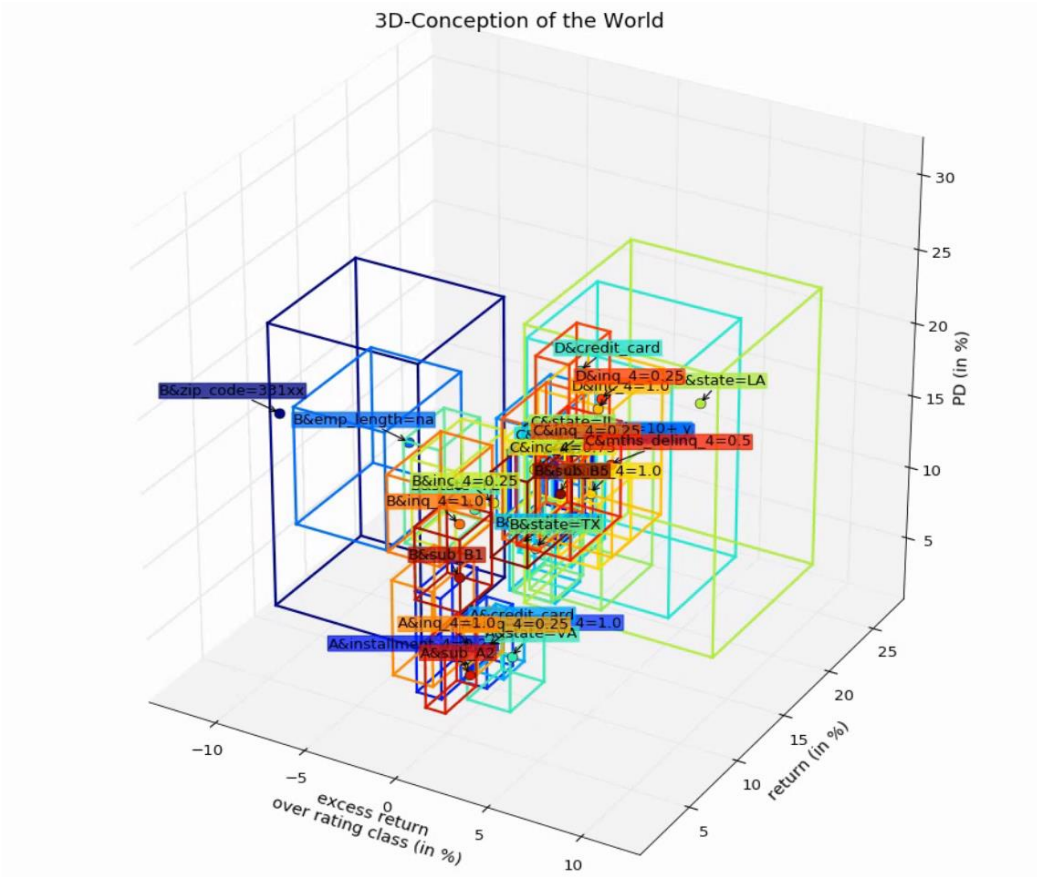
The rate of true predictions, i.e. that a pattern that was always observed up to t will appear again at $t + T$, is called **Reliability**. It was always true that the more often T-Dominance was observed up to time t (i.e. a higher DiV), the more often T-Dominance emerged again (at time $t + T$).
The fundamental finding that **Reliability** always increases with an increasing DiV was observed for several prediction and decision tasks (see: http://www.udpl.eu/udpl.html).

- Using the principle of T-Dominance is based on empirical laws – not on assumptions.
- It allows to identify the structure of underlying decision problems (see: http://udpl.eu/files/ClassifyingPredictionProblems_korr1.html).
- So the transfer of knowledge about T-Dominant strategies and their features for "similar" decision problems becomes possible.

Finally, we can use KnowledgeNets to create **World-Views** of several variables of interest. They consist of objects that had several predictable features (in the below example features are the return, the excess return of the respective rating class and the empirical rate of default).

- World-Views simply show multiple features of objects that were always true.
- If the world generates only logically consistent measurements the predictions that these features will appear at $t + T$ again are also logically consistent.
- They allow to find decision rules that are T-Dominant for multiple goals.

- The results of deductive reasoning within these predictions have always been empirically true in the past.



Graph 3: Worldview showing several always true features of selection strategies for loan portfolios (See video about the evaluation of predictions over the course of time: http://udpl.eu/learning-systems.html).

Using the above worldview an **investor** can find loan selection rules to construct a portfolio of a given size that maximizes return, considering the empirical rate of default as well as the excess return of the respective rating class at the same time. The resulting decisions take all the specified objectives into account without any logical errors between the conflicting goals.

“**Lending Club**” could use the technology to improve their rating system and define interest rates that increase the amount of loans arranged over their platform by using T-Dominant strategies. This would result in a better pricing model and a higher return for “Lending Club”.