Can a training set made from sentiment data be helpful in building a better forecasting model?

Alisa Dammer, Hamburg 24.07.2015

Structure

1. About me
2. My Bachelor thesis
3. Why you should be interested
4. What I have done already
5. What I still need to do
6. Contact

About me

Student:

- Novosibirsk State University 2007 - 2011
- University Hamburg 2011 - ...

Major Interest:

- Programming
- Econometrics, Statistics, Finance use for Strategic Planning
- Neural Networks, Machine Learning

Current Project:

- Bachelor Theme: „Sentiment Data as a fine tuning mechanism“

My Bachelor Thesis

The analysis of the **sentiment data** in the form of forecast, buzz and mood can be used as a **fine tuning** mechanism for **prediction models**. The statement will be checked for a classical econometrical model, in particular linear regression.

The idea: We will create **a linear regression** on a pre-composed training data set and check this model on a test data set. Afterwards, we will build a model with the **same variables** but on a **sentiment data** sample (SD-sample) where the SD-sample will consist of daily frequency forecasts for a 2 year period. Additionally to the new model's parameters, we will compute the mood-trend and buzz-trend. We will also check the second model on an actual test set. We will **compare the results** of both models and create a **composition model** (the parameters of the final model will computed from two previous sets of parameters). The final model will be checked on a validation set, which wasn't previously shown to any of two built models.

The hypothesis: In the **long term**, this composition of the two models will show **better results** than the two original models on their own.

Alisa Dammer, Hamburg 24.07.2015

Why should you be interested

1. On one hand, the sentiment data sets are new and not easily obtainable. This means the data is interesting for scientific research. My bachelor thesis is just one of many methods how the data can be used.
2. If the hypothesis of the thesis is correct and the sentiment data can be used as a fine tuning mechanism for simple models, the value of the data can be checked for complex algorithms and models. Positive result in further researches will make the sentiment data more interesting for practical solutions.
3. On the other hand, since the data set is highly specific, it can‘t be sold to any third party. Thus, I can guarantee that the data set will be used only for this bachelor thesis.
4. Additionally, StockPulse does not take any risks if the hypothesis of the thesis fails. The only conclusion will be that this customized data set can‘t be used with the chosen method for this particular purpose.

What I have done already

We have built a program choosing the best model with a limited number of predictors. As the result of the program we have gotten a local optimum model.

Input:

1. Pre-chosen data file in .csv format
   1. 1 learning sample
   2. 1 training sample
   3. 1 validation sample
2. First company in the table is a dependent variable

1. Statistical tests to figure out what type of model can be built
2. Set the limit on the number of the predictors
3. Step-forward approach to choose the best model
4. Determination coefficient shows the best model in the class
5. LLR-test compares two neighboor classes

Output:

1. „best match“-model
   1. Names of the predictors
   2. Estimated parameters
   3. Determination coefficient
   4. Other statistics possible (optional)

What I still need to do

To check whether the sentiment data can be useful in fine tuning of the parameters, we need a sample for **6 companies for a 2 year period, with a daily forecasts**. The time period of the data set is from **2011-01 to 2013-01**. The SD-set will consist of **forecasts**, **buzz and mood**. We will later use both buzz and mood for a proper composition.

Using the same method, we will build a new model with the same predictors that were fixed in the first step. Under the assumption, that the parameters of two models will be different, we will validate both models on the validation set. Depending on the quality of both models, we will compute a new regression and we will validate it on a completely new validation set.

Contact me

If you are intrested, you can contact me any time:

Mobile: +4917680767743

E-mail: alisa.dammer@gmail.com

Skype: d.a.alisa

For detailed information about the Bachelor Thesis, you can check out the code on GitHub: https://github.com/Alisa-lisa/uni-stuff/tree/master/Uni

For more information about me, you can check out my CV:

http://alisadammer.com/

Alisa Dammer, Hamburg 24.07.2015

Thank you for your attention and time!

Alisa Dammer, Hamburg 24.07.2015