Modern data mining in astroparticle physics

Philipp Schlunder, October 11, 2014

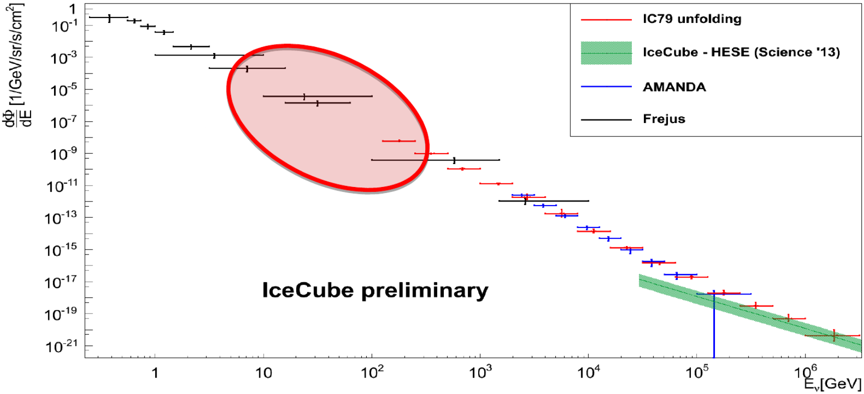
Overview

1 Motivation
- Low energy ν_μ-spectrum

2 Separation
- Feature Selection
- Multivariate Classification
- Validation

3 Conclusion

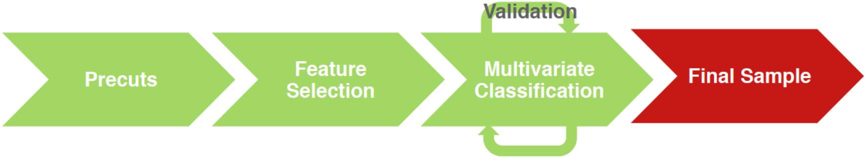
Low energy ν_μ-spectrum



Why do we need machine learning?

- Better representation of data
- Multivariate Classification
- Hidden Patterns
- Scientist: experience / knowledge / bias
  - ⇒ Precuts & feature generation

Analysis chain



Precuts

Feature Selection

Multivariate Classification

Validation

Final Sample

[1]

Feature Selection

Forward Selection [2]

1. Begin: zero attributes
2. Add one unused attribute
3. Calculate performance
4. Choose new attribute for max. performance increase
5. $\rightarrow$ 2, unless chosen number of attributes reached

Minimum Redundancy Maximal Relevance [3]

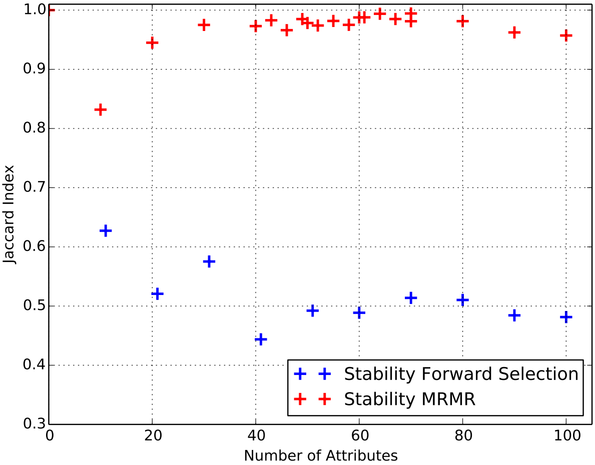
Stable represantation with min. number of attributes

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j, c) - \frac{1}{m-1} \sum_{x_j \in S_{m-1}} I(x_i, x_j) \right]$$

Stability [4]

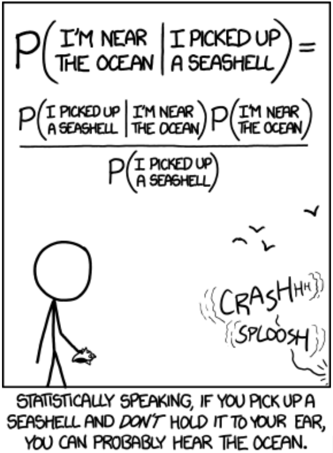
$$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|}$$

Stability Evaluation



Multivariate Classification

Naive Bayes

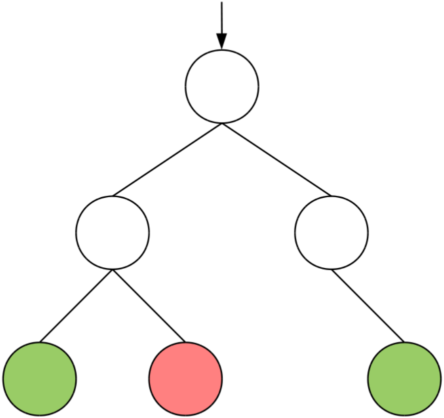

P(I'M NEAR THE OCEAN | I PICKED UP A SEASHELL) =
P(I PICKED UP A SEASHELL | I'M NEAR THE OCEAN) P(I'M NEAR THE OCEAN)
P(I PICKED UP A SEASHELL)
CRASHHH
SPLOOSH
STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND DON'T HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

[5]

$$p(C|x_1, \dots, x_n) = \frac{p(C) \prod_{i=1}^{n} p(x_i|C)}{p(x_1, \dots, x_n)}$$

Bayes Classifier [6]: $\operatorname{argmax}_c p(C = c) \prod_{i=1}^{n} p(x_i|C)$

Decision Tree



[7]

$$\text{Purity} = \frac{\sum\limits_s \omega_s}{\sum\limits_s \omega_s + \sum\limits_b \omega_b} = \frac{\text{tp}}{tp + \text{fp}}$$

$$\text{Gini} = \left(\sum_{i=1}^{n} \omega_i \right) \; P(1 - P)$$

Random Forest [2]

- Ensemble of decision trees
- Bootstrap random number of events
- Random (pre-set) number of attributes
- Less vulnerable to over fitting

$$\text{Confidence} = \frac{N_i}{N}$$

Validation

X-Validation

Train	Test	Test	Test	Test
Test	Train	Test	Test	Test
Test	Test	Train	Test	Test
Test	Test	Test	Train	Test
Test	Test	Test	Test	Train

Conclusion

Feature Selection

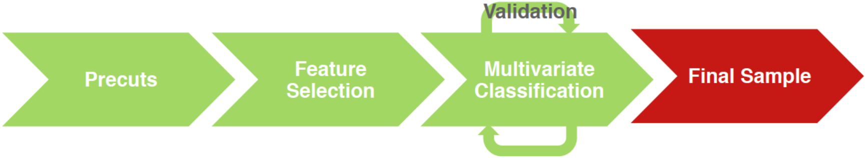
- Forward Selection
- mRMR
- Stability

Multivariate Classification

- Naive Bayes
- Decision Tree
- Random Forest

Validation

- X-Validation
- Importance



[1]

'Any measurement that you make without any knowledge of the uncertainty is meaningless.'
- Walter Lewin

References

M. Boerner. *Private Communication*. 2014.

T. Hastie, R. Tibsh, and J. Friedmann. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition 10th print. Springer, 2013.

H. Peng, F. Long, and C. Ding. “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005.

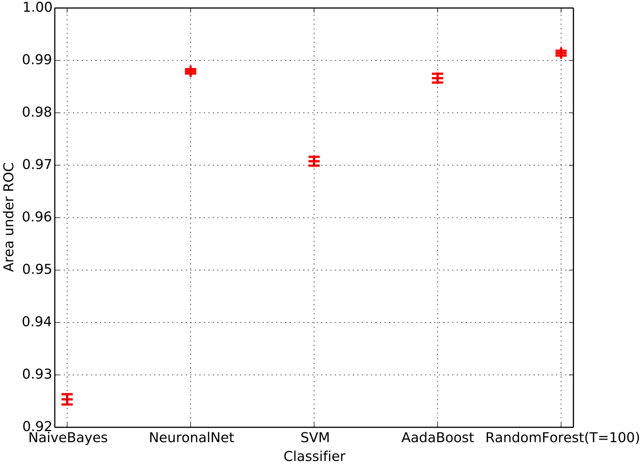
M. Levandowsky and D. Winter. “Distance between Sets”. In: *Nature* 234.5323 (Nov. 5, 1971), pp. 34–35. URL: http://dx.doi.org/10.1038/234034a0.

R. Munroe. *Seashell*. 2013. URL: http://xkcd.com/1236/ (visited on 07/15/2014).

I. Rish. *An empirical study of the naive Bayes classifier*. 2001. URL: http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf (visited on 07/15/2014).

B. P. Roe et al. “Boosted decision trees as an alternative to artificial neural networks for particle identification”. In: *Nuclear Instruments and Methods in Physics Research A* 543 (May 2005), pp. 577–584. DOI: 10.1016/j.nima.2004.12.018. eprint: physics/0408124.

Classifier Comparison


Area under ROC
NaiveBayes
NeuronalNet
SVM
AadaBoost
RandomForest(T=100)
Classifier