Filtering Knowledge: A Comparative Analysis of Information-Theoretical-Based Feature Selection Methods

Kiril Vasilev, Asterios Katsifodimos, Andra Ionescu
k.v.vasilev-1@student.tudelft.nl, {a.katsifodimos, a.ionescu-3}@tudelft.nl

Introduction 1

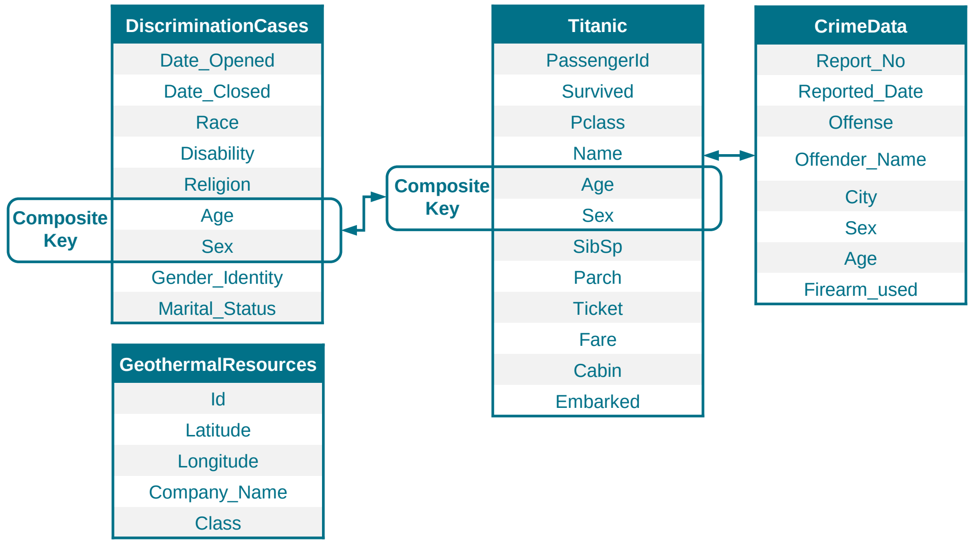

Fig. 1: An example of an augmentation scenario for "Titanic" table

Research question 2

How do the information-theoretical-based feature selection methods MIFS, MRMR, CIFE, and JMI compare in runtime and accuracy / RMSE for Machine Learning algorithms?

Information theory feature selection methods 3

Mutual Information Feature Selection (MIFS) [1]

The scoring function J for a feature X_k, a class variable Y and a set of already selected features S is as follows, where $I(X_k;Y)$ denotes the information gain between X_k and Y:

$$J_{MIFS}(X_k) = I(X_k;Y) - \beta \sum_{X_j \in S} I(X_k;X_j) \quad (1)$$

Minimum Redundancy Maximum Relevance (MRMR) [3]

$$J_{MRMR}(X_k) = I(X_k;Y) - \frac{1}{|S|} \sum_{X_j \in S} I(X_k;X_j) \quad (2)$$

Conditional Infomax Feature Extraction (CIFE) [2]

$$J_{CIFE}(X_k) = I(X_k;Y) + \sum_{X_j \in S} I(X_k;X_j|\,Y) - \sum_{X_j \in S} I(X_k;X_j) \quad (3)$$

Joint Mutual Information (JMI) [4]

$$J_{JMI}(X_k) = I(X_k;Y) + \frac{1}{|S|} \sum_{X_j \in S} I(X_k;X_j|\,Y) - \frac{1}{|S|} \sum_{X_j \in S} I(X_k;X_j) \quad (4)$$

Methodology 4

- **Datasets:** Tab. 1;
- **Algorithms:** Logistic Regression (LR), XGBoost, and SVM;
- **Metrics:** Evaluation based on runtime and accuracy / RMSE;

Tab. 1: Datasets used during evaluation

Dataset name	#Rows	#Features
Steel plates faults	1941	33
Breast cancer	569	31
Gisette	6000	5000
Internet advertisements	3279	1558
Census Income	32560	14
Housing prices	1460	80
Bike sharing	17379	16

Results (1) 5

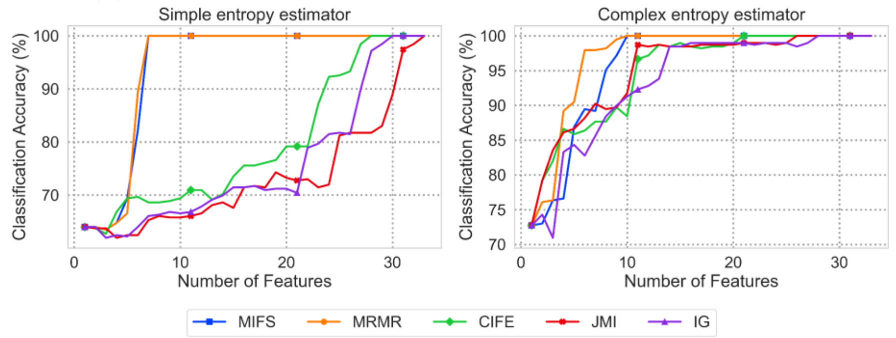

Fig. 2: Accuracy comparison of entropy estimators on Steel plates faults and LR

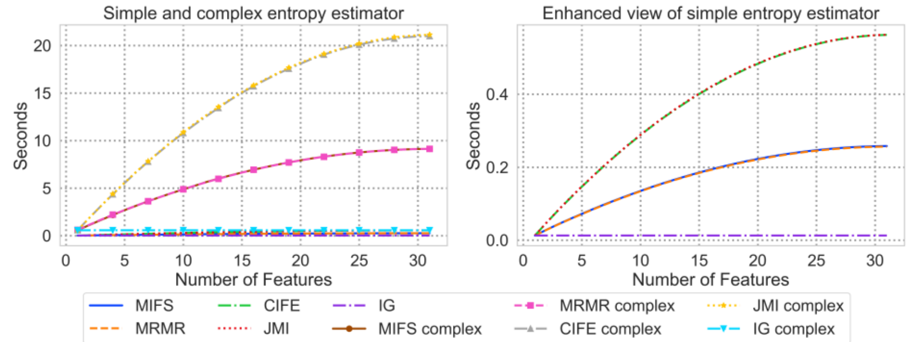

Fig. 3: Runtime comparison of entropy estimators on Breast cancer dataset

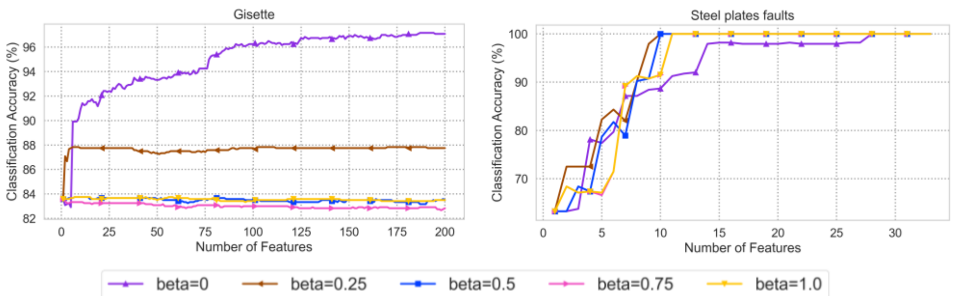

Fig. 4. Comparison of accuracy during MIFS tuning on Logistic Regression

Results (2) 6

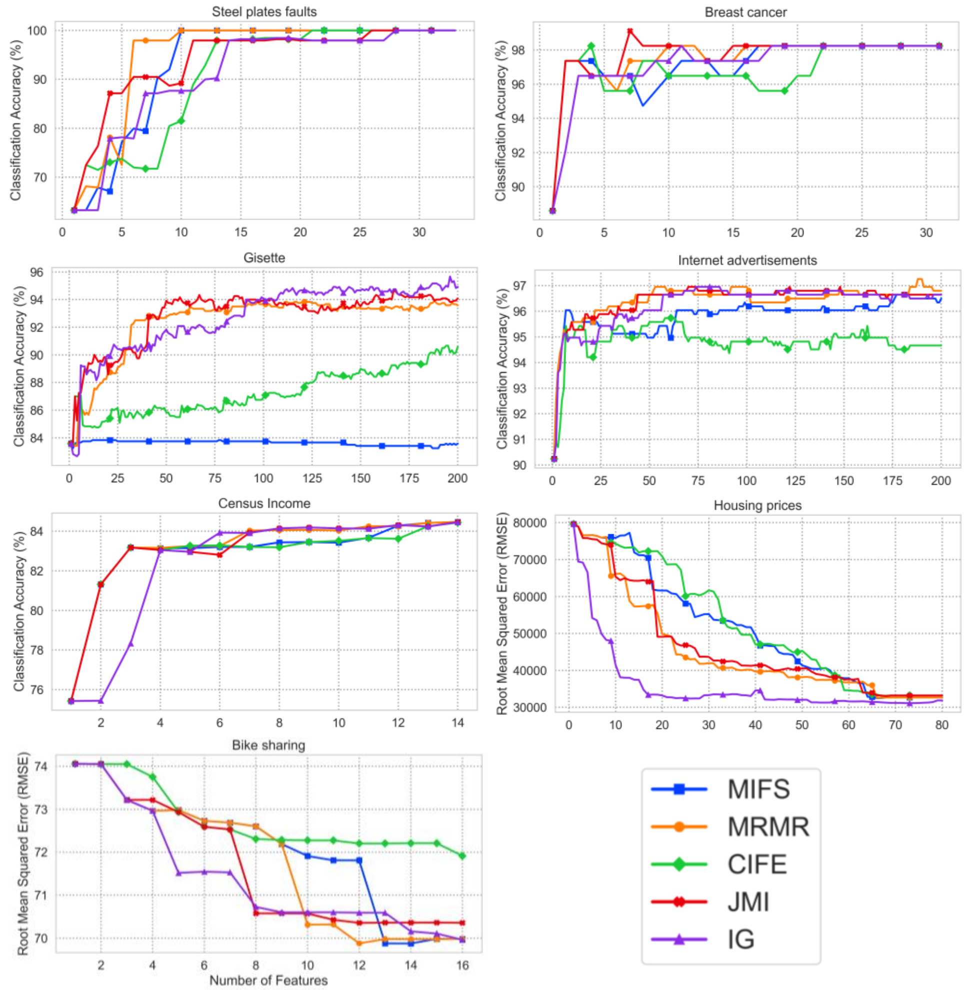

Fig. 5: Comparison of effectiveness of all datasets for Logistic Regression

Conclusions 7

- The simple entropy estimator is up to 30% less accurate, but it is 50 – 100 times quicker than the complex entropy estimator.
- MIFS and MRMR have 2 – 4 times lower runtime than CIFE and JMI.
- MRMR and JMI lead to models with significantly higher performance.
- IG feature selection can be faster and more effective than the four methods in some cases.

Limitations

- The results might be limited to the range of datasets, machine learning models used, and their hyperparameters.

References

[1] Roberto Battiti. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE trans. neural netw.* 5:537–550, 07 1994.
[2] Dahua Lin and Xiaoou Tang. Conditional infomax learning: An integrated framework for feature extraction and fusion. *ECCV*, 9:68–82, 01 2006.
[3] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-relevance, and Min-Redundancy". *IEEE TPAMI*, 27:1226–1238, 08 2005.
[4] Howard Yang and John Moody. Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. In: *Proceedings of NIPS*. Vol. 12., 1999

WIS Web Information Systems TU Delft
Web Information Systems
TU Delft – The Netherlands
https://www.wis.ewi.tudelft.nl/