Karlsruhe Institute of Technology

Institute for Information Processing Technologies

Watermelon: a Novel Feature Selection Method Based on Bayes Error Rate Estimation and a New Interpretation of Feature Relevance and Redundancy

Xiang Xie and Wilhelm Stork
xiang.xie@kit.edu, wilhelm.stork@kit.edu

Introduction

For the machine learning community, dealing with datasets which contain tens of thousands of features is not uncommon anymore. Thus, selecting a small subset of features while minimizing the generalization error is a crucial focus in such scenarios in the last decades. Currently, although the SOTA methods use different approaches to evaluate features, they generally consider the correlation between features as an indicator of pure redundancy and thus avoid selecting such features. However, we believe that only monotonic dependence is truly redundant and non-monotonic correlation may improve the performance due to their great complementarity. Based on this, we propose our method watermelon which ranks features via Bayes error rate estimation and adjusts their goodness dynamically according to their correlation.

Method

- Bayes error rate (BER) is a good measurement of the quality of features. Although It is usually unknown, we can approximate it using kernel density estimation
- We distinguish non-monotonic correlation from monotonic correlation. The former one may bring more information while the latter one is truly redundant
- The BER of a feature candidate will be penalized to maximum if there is a monotonic correlation. Two features will share their minimum if they are non-monotonic correlated
- We use an activation function to smooth the procedure and compensate the effect of weak correlation

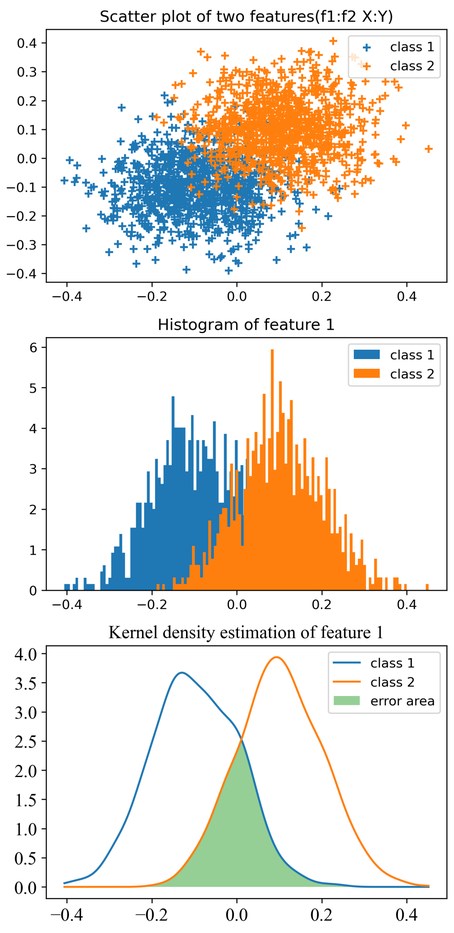

Fig. 1 Bayes error rate can be approximated using kernel density estimation

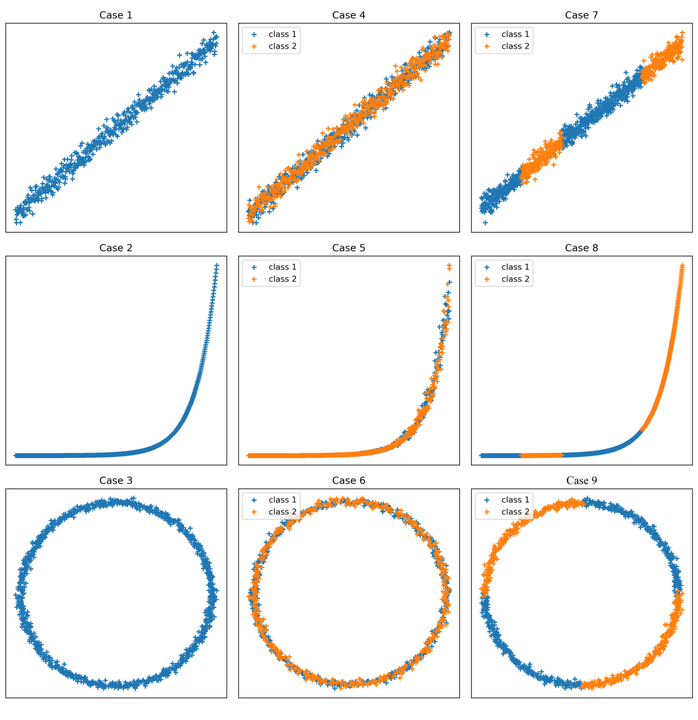

Fig. 2 Examples of two features with different types of correlation. A second feature with monotonic correlation can neither bring extra information nor improve the final performance. On the contrary, an extra feature with non-monotonic correlation may help to separate the instances under certain circumstances

Result

Method	CLL_SUB_111	COIL20	Colon	GLIOMA	Isolet	Lung	Lymphoma	nci9	ORL	orlraws10P	PCMAC	TOX_171	USPS	warpAR10P	warpPIE10P	Yale	Gisette	Avg. Rank
Watermelon	76.4	**95.9***	**84.9***	**78.0***	**86.4***	91.9	87.0	**76.4***	**90.6***	**95.8***	89.5	**85.3***	**92.1***	**93.5***	**98.9***	**74.9***	**93.7***	1.9
DISR	64.6	87.1	83.6	65.2	72.1	89.5	90.3	71.9	79.5	75.1	89.9	73.0	82.2	86.1	96.2	65.7	93.2	6.2
ICAP	63.8	80.0	81.1	63.8	70.1	89.1	**91.3***	71.9	83.7	67.8	89.7	83.8	87.0	82.3	96.6	60.8	92.5	7.2
fischer_score	58.6	79.4	79.0	77.6	74.0	88.9	86.3	73.9	81.4	80.2	88.5	79.7	86.6	84.2	97.6	66.2	92.8	7.6
JMI	62.3	79.1	83.5	64.0	69.3	90.5	91.2	70.5	78.6	91.4	89.6	67.8	88.2	84.5	97.5	61.9	92.7	7.41
f_score	58.6	79.4	79.0	77.6	74.0	88.9	86.3	74.6	81.4	80.2	88.5	79.7	86.6	84.2	97.6	66.2	82.3	7.8
trace_ratio	58.6	79.4	79.0	77.6	74.0	88.9	86.3	74.5	81.4	80.2	88.5	79.7	86.5	84.2	97.6	66.3	92.4	7.8
MRMR	55.1	89.4	79.0	68.2	77.3	86.6	91.3	70.6	82.7	72.4	**90.5***	62.2	73.3	85.0	98.3	59.3	92.2	8.0
CMIM	63.8	79.9	81.3	63.8	70.0	89.1	**91.3***	69.9	83.7	67.8	89.7	71.5	87.0	82.3	96.6	60.8	91.9	8.2
MIM	63.9	81.1	78.5	71.6	59.0	87.6	87.3	67.8	60.9	85.3	89.7	74.3	86.9	83.1	94.8	59.3	92.9	9.1
reliefF	62.5	81.0	83.2	71.4	63.2	90.4	80.2	60.1	76.8	77.8	73.4	76.8	88.0	85.8	95.5	55.9	92.8	9.1
gini_index	**79.0***	85.1	79.7	76.4	60.2	89.9	66.5	39.4	77.9	65.2	89.9	69.2	79.4	72.4	94.8	45.9	92.8	10.3
RFS	74.4	81.4	58.2	32.0	73.0	**92.7***	79.6	34.4	51.4	51.6	86.6	85.1	89.3	68.5	95.4	34.6	91.0	11.2
l1_l21	58.4	75.7	80.7	68.8	68.6	90.7	82.4	45.2	56.0	49.3	84.5	81.5	83.5	80.5	95.8	42.7	83.4	11.9
ls_l21	42.1	82.7	60.9	48.7	79.5	69.0	47.5	26.0	84.6	61.9	70.7	57.6	90.4	73.8	95.0	59.1	78.3	12.7
MIFS	54.6	55.6	78.5	48.3	64.9	84.1	85.1	46.1	79.1	77.6	86.3	55.7	72.1	63.0	95.6	46.9	84.7	13.8
FCBF	50.9	19.4	83.1	37.0	21.8	81.0	86.7	70.6	9.3	19.5	87.3	22.3	30.8	19.6	26.2	12.1	84.7	15.1
CIFE	41.7	38.7	79.8	48.0	61.4	70.7	63.3	24.3	25.4	63.5	82.0	30.0	32.7	26.2	92.8	20.3	87.6	16.0

- We compare our work with 17 SOTA algorithms on 17 datasets in different domains
- Our approach outperform others with an avg. rank of 1.9, followed by the second best method with an avg. rank of 6.2
- We use a significant dominance partial order diagram (SDPOD) to illustrate that our approach is statistically significantly better than many competitors

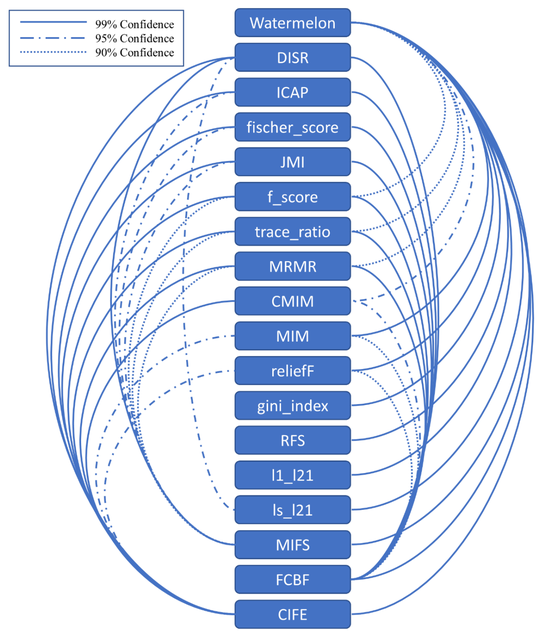

Fig. 3. SDPOD. A connection between two methods indicates a statistically significant difference. E.g. watermelon is superior to MIM with 99% confidence.

Conclusion

- Use Bayes error rate to assess features is intuitive, straightforward and effective
- Although monotonic correlation indicates true redundancy, non-monotonic relationship can improve the performance of a classifier, which is against the heuristic used by many other popular algorithms.

Reference

J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, “Feature selection: A data perspective,” ACM Comput. Surv., vol. 50, no. 6, Dec. 2017. [Online]. Available: https://doi.org/10.1145/3136625
I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of machine learning research, vol. 3, no. Mar, pp. 1157.1182, 2003.
J. R. Vergara and P. A. Est´evez, “A review of feature selection methods based on mutual information,” Neural Computing and Applications, vol. 24, no. 1, pp. 175.186, Jan. 2014. [Online]. Available: https://doi.org/10.1007/s00521-013-1368-0
P. A. Est´evez, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized mutual information feature selection,” IEEE Transactions on neural networks, vol. 20, no. 2, pp. 189.201, 2009.
G. Brown, A. Pocock, M.-J. Zhao, and M. Luj´an, “Conditional likelihood maximisation: a unifying framework for information theoretic feature selection,” Journal of machine learning research, vol. 13, no. Jan, pp. 27.66, 2012.
A. P´erez, P. Larra.naga, and I. Inza, “Bayesian classiers based on kernel density estimation: Flexible classiers,” International Journal of Approximate Reasoning, vol. 50, no. 2, pp. 341.362, 2009.