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

Why feature selection?

- Datasets with tens of thousands of features are not uncommon anymore
  - e.g. image data, microarray data and text data
- Resources & time consuming, sometimes bad performance due to noisy and/or redundant features
- What does feature selection provide:
  - Improve the prediction performance of the predictors
  - Provide faster and more cost-effective predictors
  - Better understanding of the underlying process that generated the data

Proposed method: Watermelon

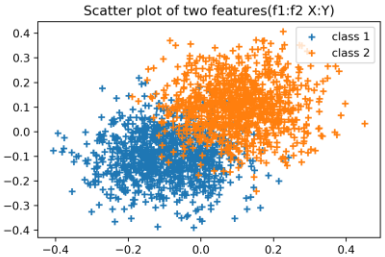
- Score each feature independently
  - Approximate Bayes Error Rate (BER)
  - Use kernel density estimation
- Greedy search of feature subset
- Dynamic adjustment of feature scores
  - Reward feature relevance
  - Penalize feature redundancy

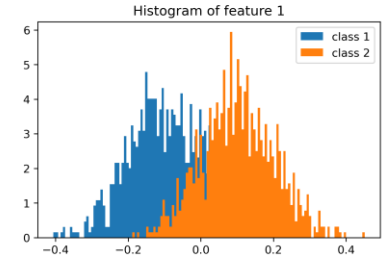
Algorithm 1 Watermelon feature selection

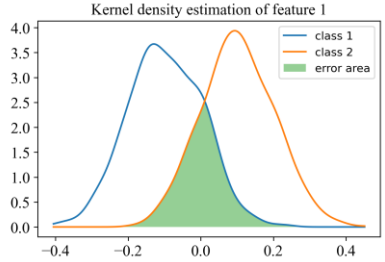
Require: $X \in R^{n \times d}$, $Y \in R^n$, L, m

init $S \leftarrow \emptyset$
for $k = 1 : d$ **do**
  for $i, j = 1 : L, i \neq j$ **do**
    calculate $BER(c_i, c_j | f_k)$
  end for
  calculate score function $Q(f_k)$
end for
$f_k^* = \underset{f_k \in F}{\operatorname{argmin}} Q(f_k), S \leftarrow S \cup f_k^*$
while $|S| < m$ **do**
  for all $f_k \in F \setminus S$ **do**
    calculate $Q(S; f_k)$
  end for
  $f_k^* = \underset{f_k \in F \setminus S}{\operatorname{argmin}} Q(S; f_k), S \leftarrow S \cup f_k^*$
  update all score functions
end while
return S

Estimation of Bayes Error Rate

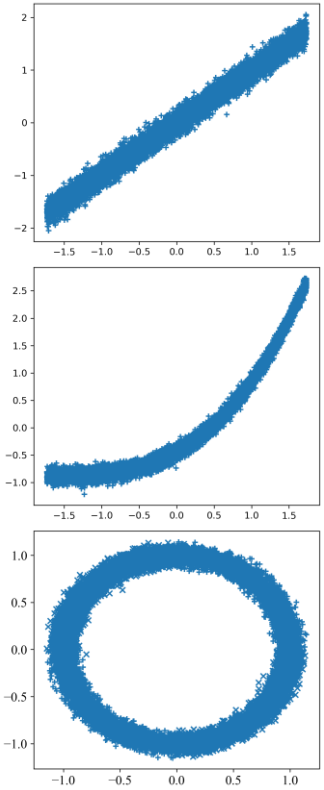






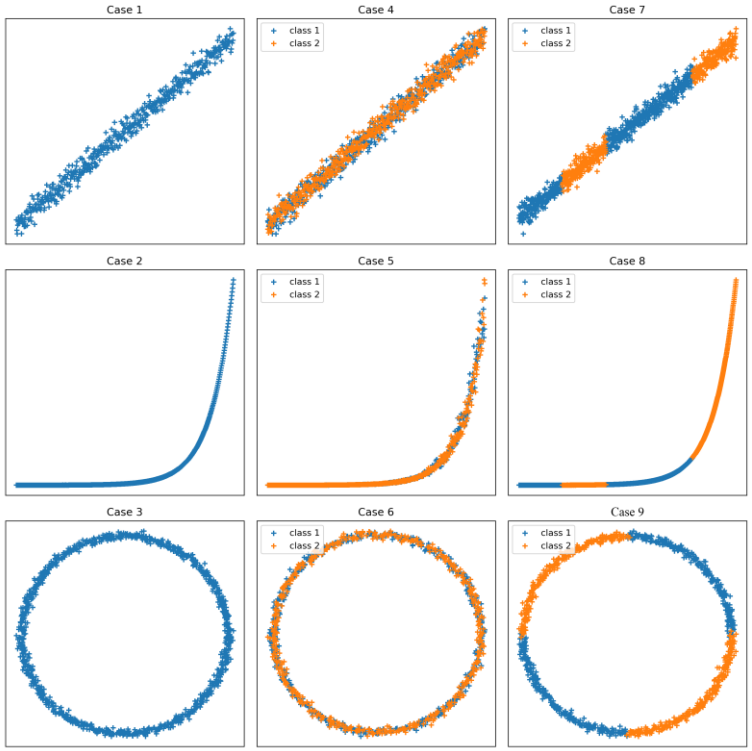
- Use kernel density estimation (KDE) to approximate the true distribution of the data
  - Because true distribution is unknown in most cases
- Calculate the **estimated** BER
  - Use one-vs-one or one-vs-all
- Score the features according to the BERs
  - Features are individually evaluated
  - ➔Assumption: features are independent

Dependence: Redundancy? Relevance?



- Measurements for dependence
  - Pearson's correlation coefficient➔ “the correlation coefficient”
    - Linear dependence
  - Spearman's rank correlation coefficient
    - Monotonic dependence
  - (Normalized) mutual information
    - General dependence

Do we need dependence?



- Monotonic dependence
  - ➔Truly redundant
  - Penalize redundancy
- Non-monotonic dependence
  - Might help due to complementarity
  - Reward relevance
- More details in paper

Comparative algorithms and datasets

17 comparative algorithms

Category	Method
similarity-based	fischer_score [4]
	reliefF [6]
	trace_ratio [5]
sparse-learning-based	ll_l21 [7]
	ls_l21 [7]
	RFS [8]
statistical-based	f_score [9]
	gini_index [10]
information-theoretical-based	CIFE [14]
	CMIM [17] [18]
	DISR [19]
	FCBF [20]
	ICAP [29]
	JMI [15] [16]
	MIFS [12]
	MIM [11]
	MRMR [13]

17 datasets

Data Type	Dataset	#instances	#features	#classes
text data	PCMAC	1943	3289	2
face image data	COIL20	1440	1024	20
	ORL	400	1024	40
	orlraws10P	100	10304	10
	warpAR10P	130	2400	10
	warpPIE10P	210	2420	10
	Yale	165	1024	15
handwritten image data	USPS	9298	256	10
	Gisette	7000	5000	2
spoken letter recognition data	Isolet	1560	617	26
biological data	CLL_SUB_111	111	11340	3
	Colon	62	2000	2
	GLIOMA	50	4434	4
	Lung	203	3312	5
	Lymphoma	96	4026	9
	nci9	60	9712	9
	TOX_171	171	5748	4

Experiment setup

- Select first n features
  - $n \in \{10, 25, 50, 75, 100, 125, 150, 175, 200\}$
- Train a linear SVM and a random forest
  - 80%/20% train/test split
  - 10 different seeds
- Fair play: no hyperparameter tuning

Results and comparison

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

Results and comparison

- Significant Dominance Partial Order Diagram
  - e.g.: Watermelon is significantly better than CIFE with 99% confidence
- Why is our method good?
- Compare with two versions
  - Watermelon-B: only use BERs to rank features
  - Watermelon-B-S: use BERs and penalize redundancy only

Classifier	Method	Avg. Acc	Avg. Rank	Overall Rank
SVM	watermelon-B	79.4	7.4	7
	watermelon-B-S	86.3	2.8	1
	watermelon	87.7	1.9	1
random forest	watermelon-B	82.0	8.8	8
	watermelon-B-S	87.3	3.7	1
	watermelon	89.7	2.7	1