Presentation:

HODE:

Hidden One-Dependence Estimator

ECSQARU 2009
on 01/07/2009

M. Julia Flores, José A. Gámez, Ana M. Martínez and José M. Puerta
Computing Systems Department
Albacete - UCLM - Spain

Outline

1. **Motivation**
2. **New proposal: Hidden One-Dependence Estimators**
   - Application of the EM algorithm
   - Number of states for the H variable
3. **Experimental methodology and results**
   - Evaluation in Terms of Accuracy
   - Evaluation in Terms of Efficiency
4. **Conclusions and Future Work**
5. **References**

Outline

1. **Motivation**

2. **New proposal: Hidden One-Dependence Estimators**
   - Application of the EM algorithm
   - Number of states for the H variable

3. **Experimental methodology and results**
   - Evaluation in Terms of Accuracy
   - Evaluation in Terms of Efficiency

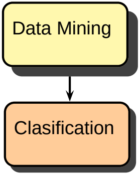
4. **Conclusions and Future Work**

5. **References**

Motivation

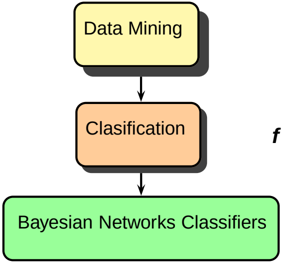
Data Mining

Motivation



$$f : X^n \longrightarrow \{c_1, \ldots, c_k\}$$

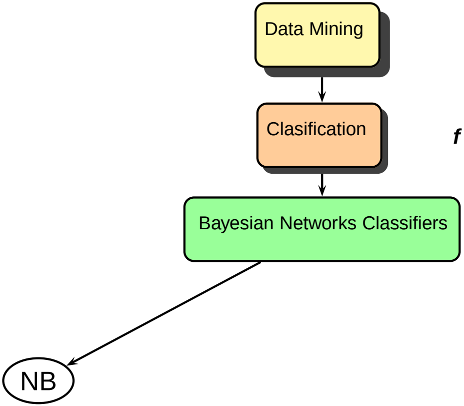
Motivation



$$f : X^n \longrightarrow \{c_1, \ldots, c_k\}$$

Bayes Theorem

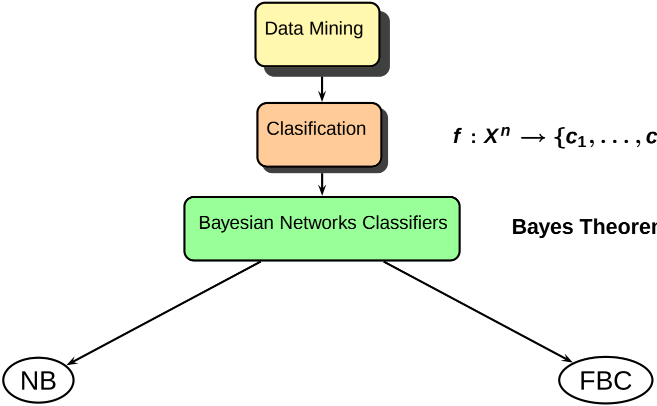
Motivation



$$f : X^n \longrightarrow \{c_1, \ldots, c_k\}$$

Bayes Theorem

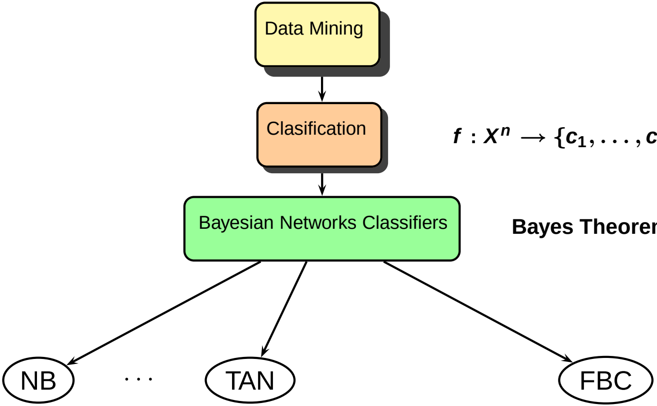
Motivation



$$f : X^n \longrightarrow \{c_1, \ldots, c_k\}$$

Bayes Theorem

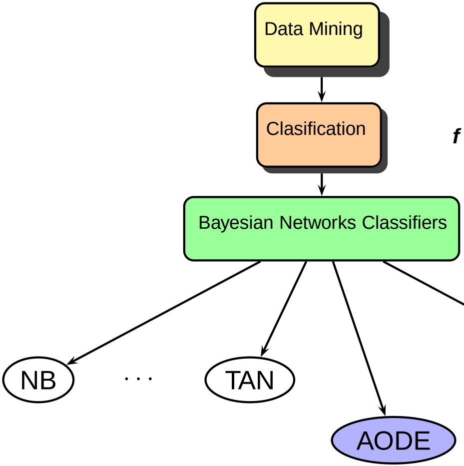
Motivation



$$f : X^n \longrightarrow \{c_1, \ldots, c_k\}$$

Bayes Theorem

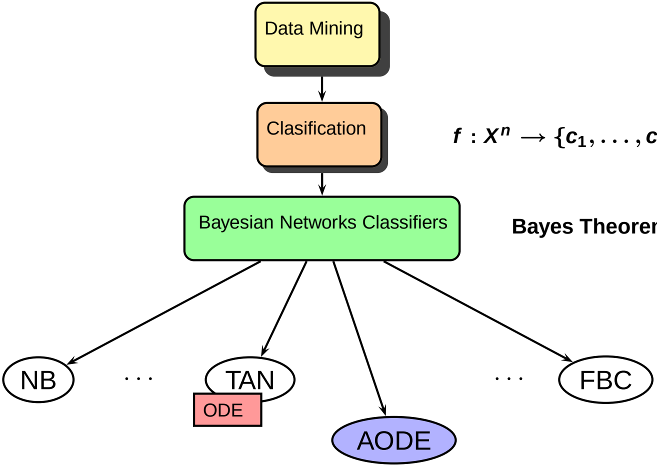
Motivation



$$f : X^n \longrightarrow \{c_1, \ldots, c_k\}$$

Bayes Theorem

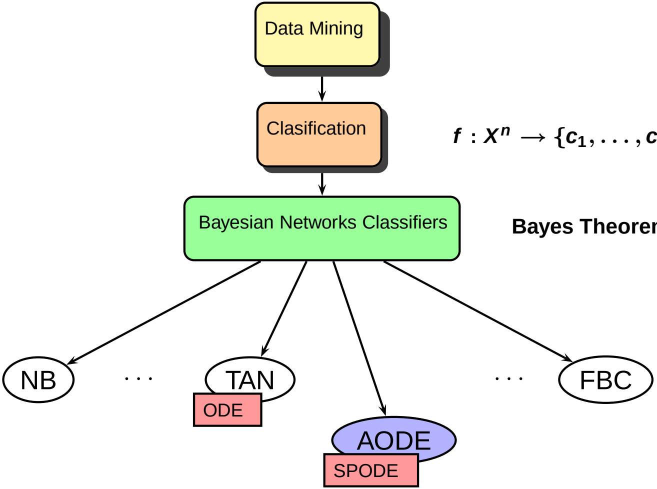
Motivation



$$f : X^n \longrightarrow \{c_1, \ldots, c_k\}$$

Bayes Theorem

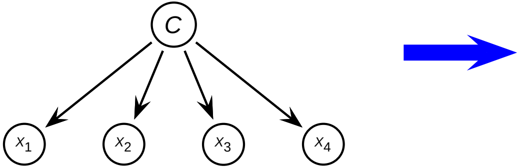
Motivation



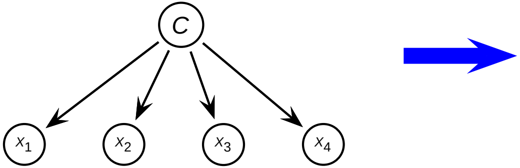
$$f : X^n \longrightarrow \{c_1, \ldots, c_k\}$$

Bayes Theorem

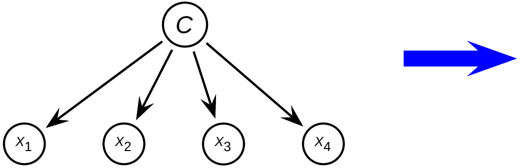
Different network structures



NB

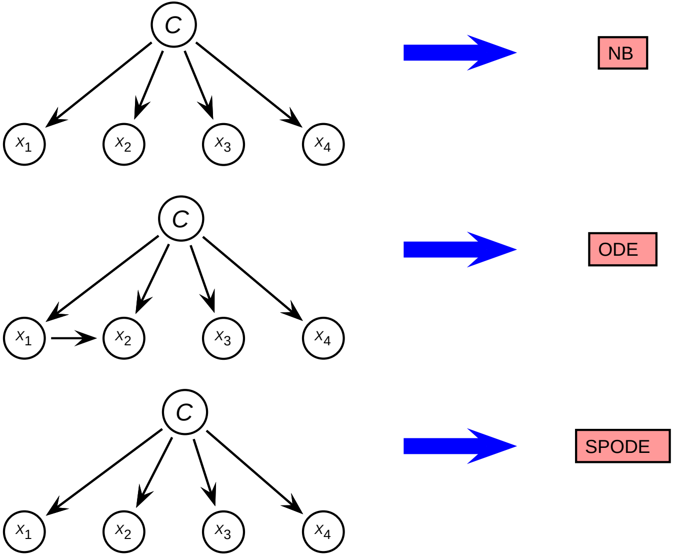


ODE

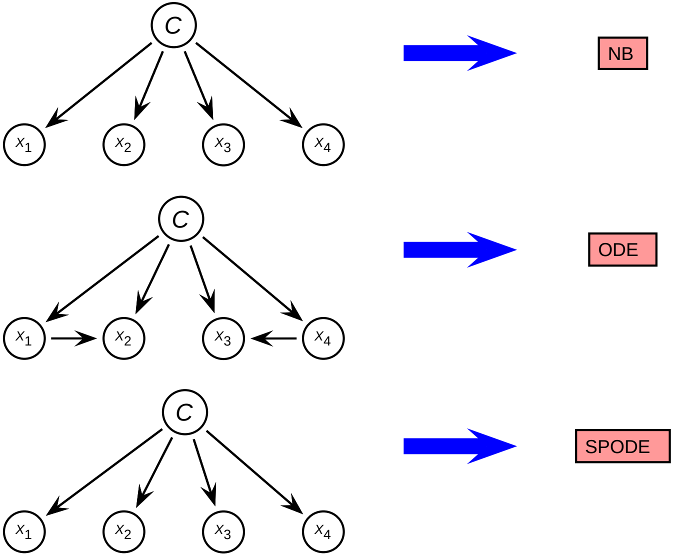


SPODE

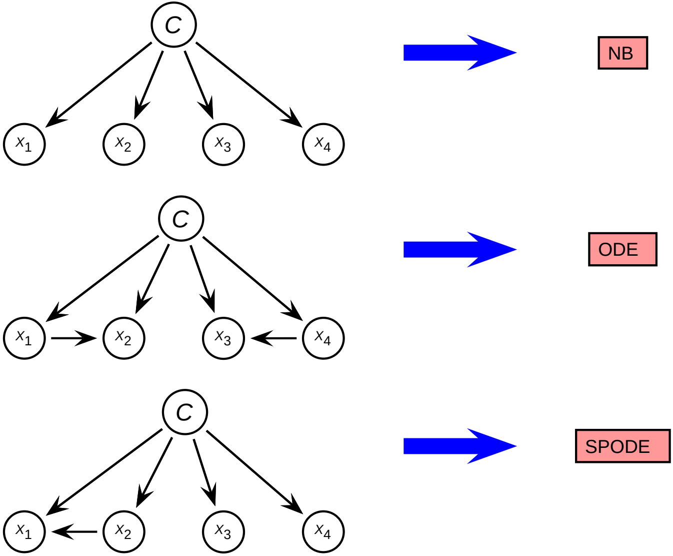
Different network structures



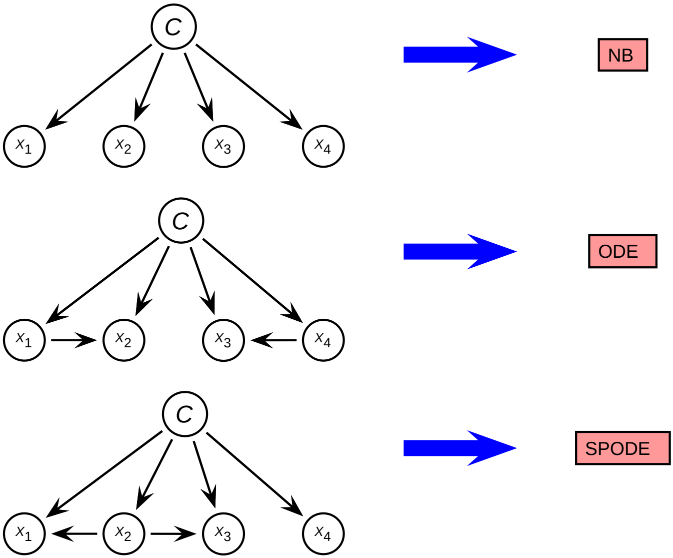
Different network structures



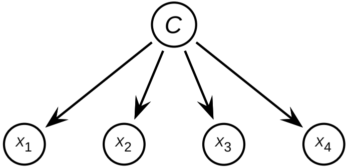
Different network structures



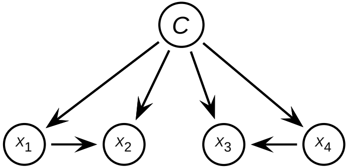
Different network structures



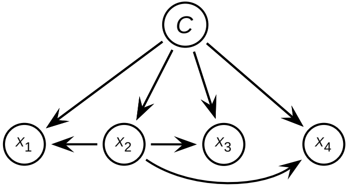
Different network structures



NB



ODE



SPODE

Naive Bayes classifier

- The attributes are conditionally independent given the class value $I(A_i, A_j|C)$.

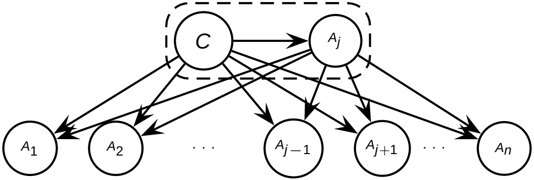
$$c_{MAP} = argmax_{c \in \Omega_C}\, p(c) \prod_{i=1}^{n} p(a_i|c)$$

- **Time complexity**:
  - **Training**: $\boldsymbol{\mathcal{O}(tn)}$
  - **Classification**: $\boldsymbol{\mathcal{O}(kn)}$

- Drawbacks:
  - × : It does not work properly in certain datasets.
  - × : Dependencies between attributes reduce, unavoidably, the prediction capability of NB.
  - × : Not only interesting to be right in the classification in certain applications.

AODE classifier I

- **AODE** is significantly better in terms of error reduction compared to the rest of semi-naive techniques.



- **MAP hypothesis:**

$$argmax_{c\in\Omega_C}\left(\sum_{j=1,N(a_j)>m}^{n} p(c,a_j)\prod_{i=1,i\neq j}^{n} p(a_i|c,a_j)\right)$$

AODE classifier II

- **Time complexity**:
  - **Training**: $\mathcal{O}(tn^2)$
  - **Classification**: $\mathcal{O}(kn^2)$

AODE classifier II

- **Time complexity**:
  - **Training**: $\mathcal{O}(tn^2)$
  - **Classification**: $\mathcal{O}(kn^2)$
- Drawbacks:

AODE classifier II

- **Time complexity**:
  - **Training**: $\mathcal{O}(tn^2)$
  - **Classification**: $\mathcal{O}(kn^2)$
- Drawbacks:
  - ✗ : **Quadratic order time in classification**.
  - ✗ : **High demand of RAM memory**.

AODE classifier II

- **Time complexity**:
  - **Training**: $\mathcal{O}(tn^2)$
  - **Classification**: $\mathcal{O}(kn^2)$
- Drawbacks:
  - ✗ : **Quadratic order time in classification**.
  - ✗ : **High demand of RAM memory**.
  - ✗ :Only discrete variables.

AODE classifier II

- **Time complexity**:
  - **Training**: $\mathcal{O}(tn^2)$
  - **Classification**: $\mathcal{O}(kn^2)$
- Drawbacks:
  - × : **Quadratic order time in classification**.
  - × : **High demand of RAM memory**.
  - × :Only discrete variables.

GAODE/HAODE ICML 09

AODE classifier II

- **Time complexity**:
  - **Training**: $\mathcal{O}(tn^2)$
  - **Classification**: $\mathcal{O}(kn^2)$
- Drawbacks:
  - × : **Quadratic order time in classification**.
  - × : **High demand of RAM memory**.
  - × :Only discrete variables.

GAODE/HAODE ICML 09

- Attempts to improve AODE's **accuracy**
  - **WAODE**: Model weighting with $MI(C, A_j)$.

Motivation

- AODE is **quadratic** in **training** and **classification time**.
  - Can be a handicap in many real applications where the **response time** is **critical**.
- The **memory required** by AODE is quite **large** due to the necessity to store the *n* models.
  - Can be an important problem when the size of the database (mainly the number of attributes) is very large.
  - Real examples: microarrays or DNA chips or KDD 09 competition.
- Our solution: new classifier which **estimates a new variable** which gathers the dependencies represented by every superparent in AODE in *a single model*.

Outline

1. Motivation

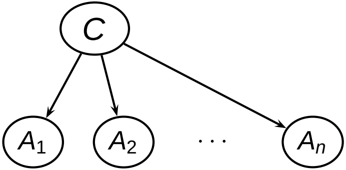
2. **New proposal: Hidden One-Dependence Estimators**
   - Application of the EM algorithm
   - Number of states for the H variable

3. Experimental methodology and results
   - Evaluation in Terms of Accuracy
   - Evaluation in Terms of Efficiency

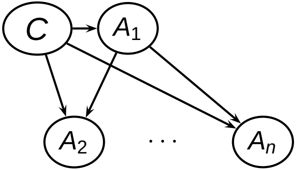
4. Conclusions and Future Work

5. References

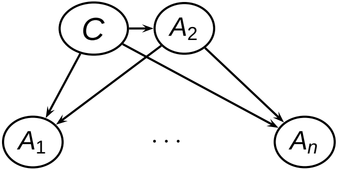
HODE Classifier



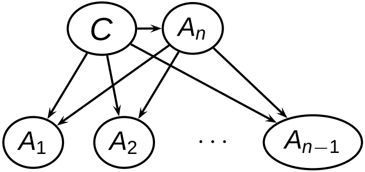
HODE Classifier



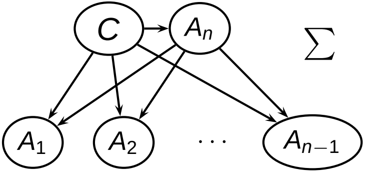
HODE Classifier



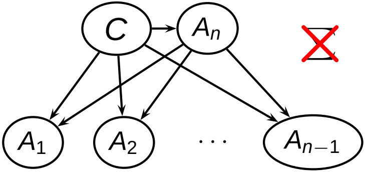
HODE Classifier



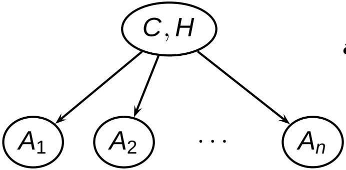
HODE Classifier



HODE Classifier

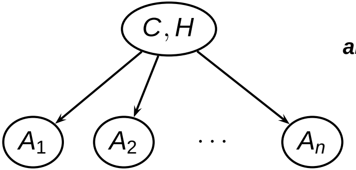


HODE Classifier



$$\operatorname{argmax}_{c \in \Omega_C} \left(\sum_{j=1}^{\#H} p(c, h_j) \prod_{i=1}^{n} p(a_i | c, h_j) \right)$$

HODE Classifier



$$\boldsymbol{argmax}_{c \in \Omega_C} \left(\sum_{j=1}^{\#H} p(c, h_j) \prod_{i=1}^{n} p(a_i | c, h_j) \right)$$

- Estimation of a new variable: **hidden variable *H***.
  - It **gathers the suitable dependencies** among the different superparents and the rest of attributes.
  - Instead of averaging the n SPODE classifiers, H's aim is to **represent the links existing in the *n* models**.
- Necessary to estimate the probability of every attribute value conditioned by the class and H.
  - **Expectation-Maximization algorithm**.

Outline

1. Motivation
2. **New proposal: Hidden One-Dependence Estimators**
   - Application of the EM algorithm
   - Number of states for the H variable
3. Experimental methodology and results
   - Evaluation in Terms of Accuracy
   - Evaluation in Terms of Efficiency
4. Conclusions and Future Work
5. References

Application of the EM algorithm

- As the different **values** for ***H*** are **not known**.
- To obtain the **maximum likelihood estimation of the parameters**.

Algorithm 1 EM algorithm adaptation to HODE.

Random initialization of weights;
{**EM ALGORITHM**}
while (!*convergence*()) **do**
    {**E-STEP**}
    Update prob. according to weights
    {**M-STEP**}
    for ($j = 1$ to $j = numInstances$) **do**
        for ($s = 0$ to $s < \#H$) **do**
            $w_{\{c,h_s,a_i,\cdots,a_n\}_j} = P(c,h_s)P(a_1|c,h_s)\cdots P(a_n|c,h_s)$;
            Normalize $\vec{w}$;
        end for
    end for
end while

Example I

Virtual dataset division

Example I

Virtual dataset division

A	B	C	H	w
a	b	c	h_1	0,3
			h_2	0,7
a	$\bar{b}$	$\bar{c}$	h_1	0,5
			h_2	0,5
$\bar{a}$	$\bar{b}$	c	h_1	0,9
			h_2	0,1
a	b	c	h_1	0,6
			h_2	0,4
$\bar{a}$	b	$\bar{c}$	h_1	0,7
			h_2	0,3
a	$\bar{b}$	c	h_1	0,2
			h_2	0,8

Example I

A	B	C	H	w
a	b	c	h_1	0,3
			h_2	0,7
a	$\bar{b}$	$\bar{c}$	h_1	0,5
			h_2	0,5
$\bar{a}$	$\bar{b}$	c	h_1	0,9
			h_2	0,1
a	b	c	h_1	0,6
			h_2	0,4
$\bar{a}$	b	$\bar{c}$	h_1	0,7
			h_2	0,3
a	$\bar{b}$	c	h_1	0,2
			h_2	0,8

Virtual dataset division

E-step

Example I

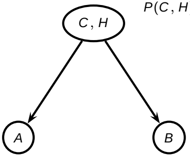
Virtual dataset division

A	B	C	H	w
a	b	c	h_1	0,3
			h_2	0,7
a	$\overline{b}$	$\overline{c}$	h_1	0,5
			h_2	0,5
$\overline{a}$	$\overline{b}$	c	h_1	0,9
			h_2	0,1
a	b	c	h_1	0,6
			h_2	0,4
$\overline{a}$	b	$\overline{c}$	h_1	0,7
			h_2	0,3
a	$\overline{b}$	c	h_1	0,2
			h_2	0,8

E-step

Structure

$P(C,H)$



$P(A|C,H)$ $P(B|C,H)$

A priori probabilities

$$p(c,h_1) = \frac{0,3+0,9+0,6+0,2}{6} = 0,33 \quad p(c,h_2) = \frac{0,7+0,1+0,4+0,8}{6} = 0,33$$

$$p(\overline{c},h_1) = \frac{0,5+0,7}{6} = 0,2 \quad p(\overline{c},h_2) = \frac{0,5+0,3}{6} = 0,13$$

CPT for attributes A and B

$$p(a|c,h_1) = \frac{0,3+0,6+0,2}{2} = 0,55 \quad p(a|\overline{c},h_1) = \frac{0,5}{1,2} = 0,42 \quad p(b|c,h_1) = 0,45 \quad p(b|\overline{c},h_1) = 0,58$$

$$p(a|c,h_2) = \frac{0,7+0,4+0,8}{2} = 0,95 \quad p(a|\overline{c},h_2) = \frac{0,5}{0,8} = 0,625 \quad p(b|c,h_2) = 0,55 \quad p(b|\overline{c},h_2) = 0,375$$

Example II

M-step

Example II

M-step

Weights count

$$p(c,h_1|a,b) = \frac{p(c,h_1)p(a|c,h_1)p(b|c,h_1)}{\sum_{i=1}^{H}(p(c,h_i)p(a|c,h_i)p(b|c,h_i))} = \frac{0{,}33 \cdot 0{,}55 \cdot 0{,}45}{0{,}254} = 0{,}32 \quad (1)$$

$$p(c,h_2|a,b) = \frac{p(c,h_2)p(a|c,h_2)p(b|c,h_2)}{\sum_{i=1}^{H}(p(c,h_i)p(a|c,h_i)p(b|c,h_i))} = \frac{0{,}33 \cdot 0{,}95 \cdot 0{,}55}{0{,}254} = 0{,}68 \quad (2)$$

Weights modification after M-step

A	B	C	H	w_1	w_2
a	b	c	h_1	0,3	0,32
			h_2	0,7	0,68
a	$\overline{b}$	$\overline{c}$	h_1	0,5	0,41
			h_2	0,5	0,59
$\overline{a}$	$\overline{b}$	c	h_1	0,9	0,92
			h_2	0,1	0,08
a	b	c	h_1	0,6	0,32
			h_2	0,4	0,68
$\overline{a}$	b	$\overline{c}$	h_1	0,7	0,79
			h_2	0,3	0,21
a	$\overline{b}$	c	h_1	0,2	0,41
			h_2	0,8	0,59

Example II

M-step

Weights count

$$p(c, h_1|a, b) = \frac{p(c, h_1)p(a|c, h_1)p(b|c, h_1)}{\sum_{i=1}^{H} (p(c, h_i)p(a|c, h_i)p(b|c, h_i))} = \frac{0{,}33 \cdot 0{,}55 \cdot 0{,}45}{0{,}254} = 0{,}32 \quad (1)$$

$$p(c, h_2|a, b) = \frac{p(c, h_2)p(a|c, h_2)p(b|c, h_2)}{\sum_{i=1}^{H} (p(c, h_i)p(a|c, h_i)p(b|c, h_i))} = \frac{0{,}33 \cdot 0{,}95 \cdot 0{,}55}{0{,}254} = 0{,}68 \quad (2)$$

Weights modification after M-step

A	B	C	H	w_1	w_2
a	b	c	h_1	0,3	0,32
			h_2	0,7	0,68
a	$\bar{b}$	$\bar{c}$	h_1	0,5	0,41
			h_2	0,5	0,59
$\bar{a}$	$\bar{b}$	c	h_1	0,9	0,92
			h_2	0,1	0,08
a	b	c	h_1	0,6	0,32
			h_2	0,4	0,68
$\bar{a}$	b	$\bar{c}$	h_1	0,7	0,79
			h_2	0,3	0,21
a	$\bar{b}$	c	h_1	0,2	0,41
			h_2	0,8	0,59

- The following **E-step** would use $\boldsymbol{w_2}$ and the cycle continues until the algorithm **converges**.
- Convergence when difference from adjacent iterations is **lower than 5 thousandths**.

Outline

Number of states for the H variable

- **Greedy technique**:
  - First step: $\#H = 1$ (NB).
  - Execution of EM algorithm and build models: $\#H = \#H + 1$.
  - Evaluate model: if worse than previous model stop process.

- **How is the fitness of the model evaluated?**
  - **Log-likelihood (LL)**:

$$LL = \sum_{i=1}^{I} \log \left(\sum_{t=1}^{\#H} p(c^i, a_1^i, \cdots, a_n^i, h_t) \right) = \sum_{i=1}^{I} \log \left(\sum_{t=1}^{\#H} p(c^i, h_t) \prod_{r=1}^{n} p(a_r^i | c^i, h_t) \right)$$

  - Penalization:
    - **Minimum Description Length**:

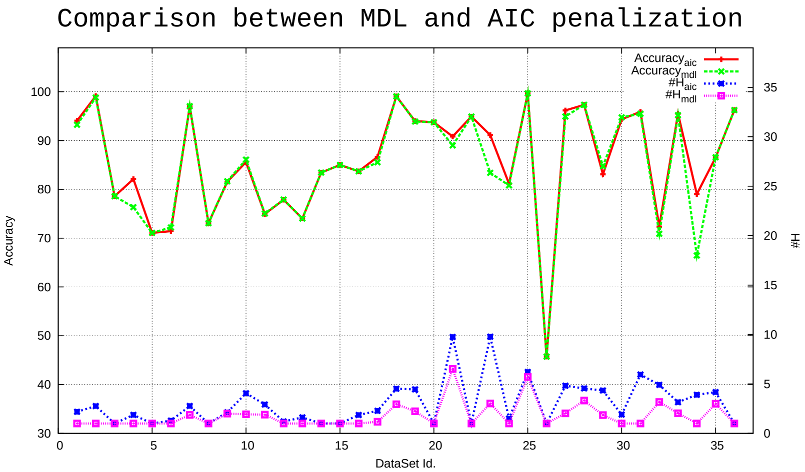
$$C(M) = \sum_{i=1}^{n} ((\#H \cdot \#C)(\#A_i - 1)) + \#H \cdot \#C - 1$$

$$MDL = LL - \frac{1}{2} \log I \cdot C(M)$$

    - **Akaike Information Criterion or AIC**:

$$AIC = LL - C(M) .$$

Accuracy and $\#H$ with AIC and MDL penalization



Outline

1. Motivation
2. New proposal: Hidden One-Dependence Estimators
   - Application of the EM algorithm
   - Number of states for the H variable
3. **Experimental methodology and results**
   - Evaluation in Terms of Accuracy
   - Evaluation in Terms of Efficiency
4. Conclusions and Future Work
5. References

Outline

Evaluation in Terms of Accuracy I

Table: Main characteristics of the datasets: number of different values of the class variable (k), number of predictive variables (n), and number of instances (I).

Id.	Dataset	k	n	I	Id.	Dataset	k	n	I
1	anneal.ORIG	6	38	898	19	ionosphere	2	34	351
2	anneal	6	38	898	20	iris	3	4	150
3	audiology	24	69	226	21	kr-vs-kp	2	36	3196
4	autos	7	25	205	22	labor	2	16	57
5	balance-scale	3	4	625	23	letter	26	16	20000
6	breast-cancer	2	9	286	24	lymph	4	18	148
7	breast-w	2	9	699	25	mushroom	2	22	8124
8	colic.ORIG	2	27	368	26	primary-tumor	21	17	339
9	colic	2	27	368	27	segment	7	19	2310
10	credit-a	2	15	690	28	sick	2	29	3772
11	credit-g	2	20	1000	29	sonar	2	60	208
12	diabetes	2	8	768	30	soybean	19	35	638
13	glass	6	10	214	31	splice	3	61	3190
14	heart-c	2	13	303	32	vehicle	4	18	846
15	heart-h	2	13	294	33	vote	2	16	435
16	heart-statlog	2	13	270	34	vowel	11	13	990
17	hepatitis	2	19	155	35	waveform-5000	3	40	5000
18	hypothyroid	4	29	3772	36	zoo	7	17	101

Evaluation in Terms of Accuracy II

Table: Accuracy results obtained with AODE and HODE classifiers.

Dataset	AODE	HODE	$\overline{\#H}$	Dataset	AODE	HODE	$\overline{\#H}$
anneal.ORIG	93,3185	**•94,0646**	2, 2	ionosphere	92,9915	**•93,9886**	4, 4
anneal	98,196	**•99,1203**	2, 8	iris	93,2	**•93,7333**	1
audiology	71,6372	**•78,5841**	1	kr-vs-kp	**91,0325**	90,8229	9, 7
autos	81,3658	**•82,0975**	1, 9	labor	**95,0877**	94,9123	1
balance-scale	69,344	**•71,088**	1	letter	88,902	**•91,117**	9, 8
breast-cancer	**•72,7273**	71,4336	1, 3	lymph	**•87,5**	81,1487	1, 5
breast-w	96,9671	**96,9814**	2, 8	mushroom	**•99,9508**	99,6824	6, 2
colic.ORIG	**•75,9511**	73,0707	1	primary-tumor	**•47,8761**	45,7227	1
colic	**•82,5543**	81,5489	2, 1	segment	95,7792	**•96,1732**	4, 8
credit-a	**•86,5507**	85,5942	4, 1	sick	**•97,3966**	97,3118	4, 6
credit-g	**•76,33**	74,94	2, 9	sonar	**•86,5865**	83,0769	4, 3
diabetes	**•78,2292**	77,8516	1, 2	soybean	93,3089	**•94,3631**	1, 9
glass	**•76,2617**	74,0187	1, 6	splice	**•96,116**	95,8872	3, 9
heart-c	83,2013	**•83,4323**	1	vehicle	72,3049	**72,3522**	4, 9
heart-h	84,4898	**85,0**	1	vote	94,5288	**•95,5173**	3, 1
heart-statlog	82,7037	**•83,7037**	1, 9	vowel	**•80,8788**	79,0101	3, 9
hepatitis	85,4839	**•86,6452**	2, 3	waveform-5000	86,454	**86,54**	4, 2
hypothyroid	98,7513	**•99,0668**	4, 5	zoo	94,6535	**•96,2376**	1

Evaluation in Terms of Accuracy III

- **Summary with AIC: ($w/t/l$, using a two-tailed t-test)**
  - $16/6/14$ (with a 95 % confidence level).
  - $15/8/13$ (with a 99 % confidence level).
- No significant difference with Wilcoxon test.
- **Summary with MDL:**
  - $11/7/18$.

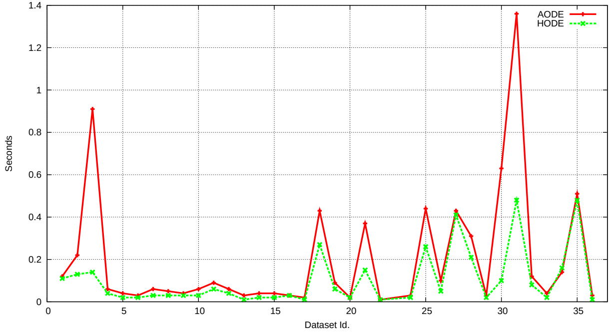
Outline

1. Motivation
2. New proposal: Hidden One-Dependence Estimators
   - Application of the EM algorithm
   - Number of states for the H variable
3. **Experimental methodology and results**
   - Evaluation in Terms of Accuracy
   - **Evaluation in Terms of Efficiency**
4. Conclusions and Future Work
5. References

What could make us vote for one or the other?

Time complexity

- At **training time** HODE is **quadratic in the worst case**:
  - $1tn + 2tn + \cdots + ntn$
  - AODE is usually faster in model construction (because of EM).
- At **classification time** HODE is **linear**, whereas AODE's is quadratic.



What could make us vote for one or the other?

Space complexity

- **Lower than AODE's**, it needs to store less CPTs.
- $\mathcal{O}(kn\#Hv)$ for HODE vs $\mathcal{O}(k(nv)^2)$ for AODE.
- AODE demands more RAM memory, problems in large databases with a high number of attributes or even attributes with lots of states.
- Experiments in 7 datasets of microarrays or DNA chips.

Dataset	k	n	l	NB	AODE	HODE
colon	2	2000	62	93,5484	91,9355	**96,7742**
DLBCL-Stanford	2	4026	47	100	100	100
GCM	14	16063	190	60,5263	OutOfMem	**70**
leukemia	2	7129	72	**100**	OutOfMem	98,6111
lungCancerHarvard2	2	12533	181	98,895	OutOfMem	**99,4475**
lymphoma	9	4026	96	**96,875**	OutOfMem	75
prostate_tumorVS	2	12600	136	80,1471	OutOfMem	**95,5882**

- OutOfMem: problems of overflow with a maximum of 8 gigabytes.
- HODE terminated without problems, even with a lower need for memory.

Outline

Conclusions and Future Work

- **HODE**: alternative to the AODE classifier:
  - Results in terms of accuracy similar to AODE.
  - **Linear order** in *classification time*.
    - Lower time response in many real applications.
  - **Reduction** in *space complexity*.
    - Lower RAM consumption.
- HODE tested in a **parallel environment**: global optimum for $\#H$.
- Additional improvements:
  - Direct adaptation for the imputation of **missing values** in the dataset, use of **EM**.
  - Average the constructed models and more.
- Clear alternative to AODE in many real applications: KDD 09.

Thank you

Outline

References I

- Zheng, F., Webb, G.I.: A Comparative Study of Semi-naive Bayes Methods in Classification Learning. In: 4th Australasian Data Mining Conference (AusDM05), Simoff, S.J, Williams, G.J, Galloway, J., Kolyshkina, I. editors, pp. 141–156. University of Technology, Sydney (2005)
- Flores, M. J., Gámez, J. A., Martínez, A. M. and Puerta J. M.: GAODE and HAODE: Two Proposals based on AODE to Deal with Continuous Variables. In: 26th International Conference on Machine Learning: 40. Montreal, Canada (2009)
- Webb, G. I., Boughton, J. R., Wang, Z.: Not So Naive Bayes: Aggregating One-Dependence Estimators. J. Mach. Learn. 58 (1), 5-24 (2005)
- Lowd, D., Domingos, P.: Naive Bayes models for probability estimation. In: 22nd international conference on Machine learning, pp. 529–536. ACM, Bonn (2005)

References II

- Cheeseman, P., Stutz, J.: Bayesian classification (AutoClass): theory and results. Advances in knowledge discovery and data mining. pp. 153–180. AAAI Press (1996)