Desalination and Water Treatment
www.deswater.com
21 (2010) 102–108
September
1944-3994 / 1944-3986
doi: 10.5004/dwt.2010.1265

Fault diagnosis for an MSF desalination plant by using Bayesian networks

Enrique E. Tarifa[a]*, F. Núñez Álvaro[a], Samuel Franco[a], Sergio Mussati[b]

[a]*Universidad Nacional de Jujuy - CONICET, Gorriti 237, 4600 San Salvador de Jujuy, Argentina*
Tel. +54 3884221587; Fax +54 3884221581; email: eetarifa@arnet.com.ar
[b]*INGAR – CONICET, Avellaneda 3657, 3000 Santa Fe, Argentina*

Received 30 September 2008; Accepted 7 April 2010

ABSTRACT

This work outlines the development of a fault diagnostic system for an MSF (multi-stage flash) desalination plant by using BNs (Bayesian networks). This diagnostic system processes the plant data to determine whether the process state is normal or not. In the latter case, the diagnostic system determines the cause of the abnormal process state; i.e., it finds out which is the fault that is affecting the supervised process. A BN is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. A BN readily handles situations where some data entries are missing. This paper determines both the structure and parameters of a BN intended for a diagnostic system. The implemented system is evaluated by using a dynamic simulator, which was developed for a real MSF desalination plant. Besides, the diagnostic system performance is compared with the performances of two other diagnostic systems. The obtained results show some advantages for the BN based diagnostic system.

Keywords: Fault diagnosis; Bayesian networks; Dynamic simulation; MSF desalination plant

1. Introduction

Quick and correct detection and identification of process faults (i.e., fault diagnosis) are extremely important when efficient, economic and safe operation of chemical processes is concerned. Undetected process fault may lead to poor quality off-spec products, resulting in poor plant economy and sometimes even catastrophic consequences like accidents, injury to plant personnel. Successful detection and identification of process faults at an early stage can increase the rate of fault recovery during operations, preventing in this manner costly accidents and unnecessary shutdowns. However, those tasks are difficult for operators of industrial plants. Indeed, there are too many process variables to be continuously supervised, and the relation among those variables and the potential faults usually is rather complex. For all those reasons, diagnostic systems are developed to aid the operators to detect and indentify faults.

* Corresponding author.

Several methodologies have been proposed for fault detection and identification in chemical processes. Those methodologies can be classified in the following groups: quantitative model-based methods [1], qualitative models and search strategies [2] and process history based methods [3]. Each method has it own strengths and weaknesses for practical applications; therefore, they can be judiciously combined to yield a better system [4,5].

BNs (Bayesian networks) have been successfully applied in fault diagnosis [6,7]. A BN is used to model a domain containing uncertainty in some manner. This uncertainty can be due to imperfect understanding of the domain, incomplete knowledge of the state of the domain, randomness in the mechanisms governing the behavior of the domain, or a combination of those. A BN is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. A BN readily handles situations where some data entries are missing. It is also an ideal representation for combining prior knowledge (which often comes in causal form) and data because the model has both a causal and probabilistic semantics. Bayesian statistical methods in conjunction with BNs offer an efficient approach for avoiding the overfitting of data [8].

This work outlines the development of a fault diagnostic system for a MSF (multi-stage flash) desalination plant by using BNs. A diagnostic system processes the plant measured data to determine whether the process state is normal or not. In the latter case, the diagnostic system determines the cause of the abnormal state; e.g., a damaged piece of equipment, an operator's action, an alteration of the process inputs. In this work, that cause is considered a fault [9,10].

When a fault affects a plant, a process parameter or variable is directly perturbed by it; then, that original perturbation propagates itself throughout the plant taking the process variables away from their normal values. The evolution of the process variables is a function of the process and fault characteristics. Indeed, there will be different evolutions depending on which process parameter or variable is directly affected by the fault and the form of the original perturbation. In this work, the form of that original perturbation is established by two parameters: the magnitude and the development period. The former specifies the maximum magnitude by which the normal value of the directly affected process parameter or variable will be perturbed by the fault; the latter specifies the time elapsed from the perturbation beginning, with null magnitude, and the perturbation full development, when it reaches the maximum magnitude. Therefore, serious faults are modeled with large magnitudes; whereas mild faults are modeled with small magnitudes; abrupt faults are modeled with short development periods; whereas gradual faults are modeled with long development periods.

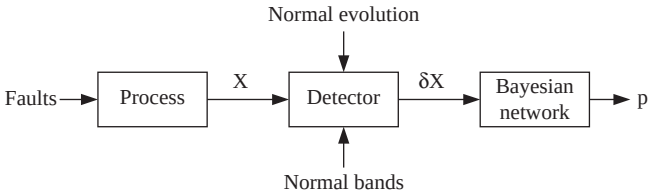

Fig. 1. Fault diagnostic system structure.

Fig. 1 shows the structure of the diagnostic system proposed in this work. For each sampling period, the plant measured data are processed by a detector block whose mission is to compare the current process variables values X with the values Xn corresponding to the normal evolution. The detector output is a vector of standardized deviations δX. This vector is sent to a BN. The network output is a vector of probabilities p, therefore its components are values between 0 and 1. Every component of p is the probability associated to a potential fault. The fault with the highest probability is the most probable cause of the abnormal state reported by the detector block.

In this work, a real desalination plant is studied. Tarifa and Scenna [11] developed the dynamic simulator MSF 2000 for the considered plant. That simulator was used in this work to evaluate de proposed diagnostic system.

2. Faults and sensors

The first step in the development of the diagnostic system is the determination of the scope of application. This scope is given by the list of faults that must be recognized by the diagnostic system and the desired format of the report. In this work, except the ''Recovery low'' and ''Rejection low'', all the faults modelled in the simulator MSF 2000 [11] are considered: a total of 55 faults and the normal state.

The desired format of the report is the probability, a number belonging to the closed interval [0, 1], for each potential fault. This number represents the certainty about the corresponding fault is affecting the plant. The higher is the probability, the higher is the certainty of the affirmation.

For a given scope of application, not all variables provide useful information. Irrelevant data may complicate the structure and affect the performance of the diagnostic system. To identify the relevant variables, the outputs of the simulator MSF 2000 [11] must be

Table 1
Normal bands

2°C for temperatures.
500 tn/h for flow rates.
5 cm for levels.
2% for outputs of controllers.
for Rmus (set point of make-up controller).

carefully analyzed. Once identified those variables, only the corresponding sensors must be selected to be supervised by the diagnostic system. To identify the relevant variables, all the potential faults were simulated for a set of magnitudes and development periods. The evolution obtained for a given fault, magnitude and development period is defined as a dynamic state. All the dynamic states obtained in that way were then analyzed to select the suitable sensors by applying the rules suggested by Tarifa and Scenna [9]. From the analysis of 270 dynamic states and the normal half bands showed in Table 1, 30 sensors were selected; they are shown in Table 2.

3. Detector block

The mission of the detector block is to detect any dynamic state as soon as possible. To do that, it calculates standardized deviations from the process data, the normal evolution and the normal bands. The standardized deviation for the variable j is defined as:

$$\delta X_j = \frac{X_j - Xn_j}{\Delta Xn_j}, \tag{1}$$

where X_j is the variable value, Xn_j is the normal value (for steady state, it is constant) and $\triangle Xn_j$ is the normal half band for the considered variable. The value of a given variable is classified as normal if it belongs to the open interval $(Xn_j - \triangle Xn_j,\ Xn_j + \triangle Xn_j)$. The normal half band is fixed taking in to account the disturbances (i.e., noises) that affect the variable during the normal operation of the process. The advantage of using standardized deviations is they can be compared with one another (they are dimensionless numbers). Moreover, a dynamic state can be detected when the absolute value of a standardized deviation becomes equal or higher than 1, that time is the detecting time t_0.

For a given sampling period $\triangle t$, the sampling time t_k is defined as:

$$t_k = t_0 + k\,\Delta t. \tag{2}$$

Since a qualitative method was selected for this work, and the ''first change'' approach [9] is used to process the data, the standardized deviations are transformed into qualitative ones according to:

$$\delta X_{j,k} \leftarrow \begin{cases} \delta X_{j,k-1}\ (k>0) \wedge \left(\left|\delta X_{j,k-1}\right| = 1\right) \\ 1 \qquad\qquad \delta X_{j,k} \geq 1 \\ -1 \qquad\qquad \delta X_{j,k} \leq -1 \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{3}$$

In this way, qualitative deviations can only be 0, −1 or +1. In the normal state, all qualitative deviations are null. Once a qualitative deviation becomes −1 or +1, it keeps that value.

4. Bayesian networks

4.1. Definition

As was stated above, when a fault occurs, it originates a sequence of symptoms; however, some

Table 2
Selected sensors

Controller CL	Controller CH	Controller CT	Controller CW
Lm (cm)	T0m (°C)	Tcwm (°C)	Wcwm (tn/h)
Ls (cm)	T0s (°C)	Tcws (°C)	Wcws (tn/h)
EL (cm)	ET0 (°C)	ETcw (°C)	EWcw (tn/h)
AL (pc)	AT0 (pc)	ATcw (pc)	AWcw (pc)

Controller CM	Controller CR	Others
Wpm (tn/h)	Wbm (tn/h)	L[17] (cm)
Wmum (tn/h)	Wbs (tn/h)	Pvh (atm)
Rmus (frac)	EWb (tn/h)	T[1] (°C)
EWmu (tn/h)	AWb (pc)	Wbd (tn/h)
AWmu (pc)		Whw (tn/h).

symptoms may be detected in an unexpected sequence (i.e., sooner or later than when they are expected), or not be detected at all. The reason for that difficulty is the size of the normal half bands, which is adjusted according to the signal-to-noise ratio of each variable. Therefore, the diagnostic system must be able to handle the uncertainty associated to the information provided by the detector block. BNs, by definition, possess such ability.

A BN is a network of nodes connected by directed links with a probability function attached to each node [6–8]. The network of a BN is a DAG (directed acyclic graph); that is, there is no directed path starting and ending at the same node. A node represents either a discrete random variable with a finite number of states or a continuous (Gaussian distributed) random variable. A link between two nodes represents causal relationships between them. In this work, only discrete variables were considered.

If a node does not have any parents (i.e., there are not nodes with links pointing towards it), the node will contain a marginal probability table (also called unconditional probability table). If a node does have parents (i.e., there is one or more links pointing towards it), the node contains a CPT (conditional probability table). If the node is discrete, each cell in its CPT contains a conditional probability for the node being in a specific state given a specific configuration of states of its parents. Thus, the number of cells in a CPT for a discrete node equals the product of the number of possible states for the node and the product of the number of possible states for the parent nodes.

The building of a BN involves two stages: the drawing of the DAG and the determination of the CPT of every node. Both stages can be carried out by using expert knowledge only. However, there are works aimed to obtain automatically both DAG and CPT from available data by utilizing structure learning (i.e., the task of drawing the DAG from data) and parameter learning (i.e., the task of determining the CPT from data) [8]. This work use a mixed approach: the DAG is obtained from a traditional BN type, whereas the parameter learning is automatic from data.

Once completed the BN, it calculates the probability associated to every node. When new data or evidences are available, the BN updates those probabilities by inference, which means computing the conditional probability for some variables given information (evidence) on other variables. This is straightforward when all available evidence is on variables that are ancestors of the variables of interest. But when evidence is available on a descendant of the variables of interest (which is the case in this work), the BN has to perform inference opposite the direction of the edges. To this end, Bayes' Theorem is employed:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}, \tag{4}$$

where $P(A \mid B)$ is the conditional probability of a node being on state A assuming that a descendant is on state B.

4.2. Classification

In this work, fault diagnosis is considered as a classification problem; BNs are often used for that kind of problems. In the classification learning problem, a learner attempts to construct a classier from a given set of labeled training examples that are represented by a tuple of attribute variables used collectively to predict the value of the class variable [12]. In the fault diagnosis problem, the attribute variables are the qualitative deviations provided by the detector block, and the class variable represents the potential faults.

The Naive Bayes is the simplest BN classifier [12], in which each attribute node (corresponding to an attribute variable) has the class node (corresponding to the class variable) as its parent, and does not have any other parent. Fig. 2 shows the Naive Bayes classifier implemented in this work. Since the class node has 56 possible states (55 faults and the normal state) and 30 sensors were selected – each of them with three possible states: -1, 0 or $+1$ –, the CPT belonging to each attribute node has 168 conditional probabilities. Therefore, a total of 5040 conditional probabilities have to be determined. Those values were obtained by processing 1100 dynamic states, and by evaluating the relative frequency associated to each cell of every CPT. The dynamic states were obtained by using the simulator MSF 2000 [11]; for each potential fault, 20 dynamic states were simulated for the combination of 25%, 50%, 75% and 100% of magnitude with 0%, 25%, 50%, 75% and 100% of development period. The unconditional probabilities associated to each state of the class node were obtained by assuming a uniform distribution; therefore, each probability is equal to 1/56; a

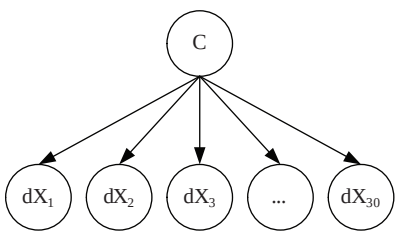

Fig. 2. Naive Bayes classifier.

Table 3
Qualitative deviations for fault #CH Set Point High

t (min)	T0s	AT0
0	0	0
4.24	0	1
9.09	1	1

better alternative is to estimate those unconditional probabilities from maintenance department records.

The Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions; namely, the Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. In spite of their naive design and apparently over-simplified assumptions, Naive Bayes classifiers often work well in many complex real-world situations [13]. Zhang [14] carried out a careful analysis of the Bayesian classification problem, and showed that there are some theoretical reasons for the apparently unreasonable efficacy of Naive Bayes classifiers.

4.3. System evaluation

To evaluate the proposed diagnostic system, every potential fault was simulated by using the simulator MSF 2000 [11] for a set of magnitudes and development periods different from those employed in the learning stage. The outputs were transformed into qualitative deviations, and they were entered into the Naive Bayes classifier. The network was implemented with the software GeNIe 2.0.

Tables 3 and 4 present the evolution of the first detected symptoms for the faults #CH Set Point High (set point of controller CH becomes higher than the normal) and #CH Set Point Low (set point of controller CH becomes lower than normal), respectively. For both cases, the fault was simulated beginning at 0 min with 55% of magnitude and 45% of development period. The detecting time t_0 was equal to 4.24 min, at that time the detector block detected the first symptom and started the classifier up. Figs. 3 and 4 contain the

Table 4
Qualitative deviations for fault #CH Set Point Low

t (min)	T0s	AT0
0	0	0
4.24	0	−1
9.09	−1	−1

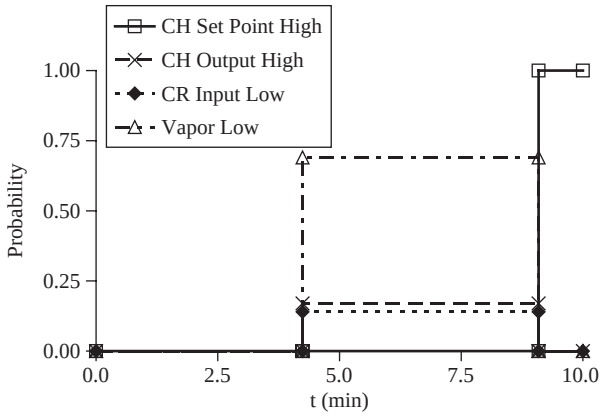

Fig. 3. Diagnostic for fault #CH Set Point High.

corresponding classifier outputs. Considering those figures, the diagnostics are correct for both cases; i.e., only the simulated fault was placed at the top with probability equal to 1 almost instantaneously. A similar behavior was observed for all the tested cases. In general, the faults more serious and abrupt (they have large magnitudes and short development periods), are more quickly identified by de classifier. That is a convenient feature.

If the magnitude decreases and the development period increases, classifier efficacy decreases. In fact, for faults simulated with 50% of magnitude and 50% of development period, the classifier efficiency degrades to 70% (rejecting probabilities below 0.5). That is not due to an intrinsic limitation of the Naive Bayes classifier; rather, that occurs because, for those magnitude and development period values, the faults are not strong enough to originate the number of symptoms needed to identify them. For example, Table 5 shows the symptoms detected when the fault #CH Output Low (output of controller CH is lower than normal) with

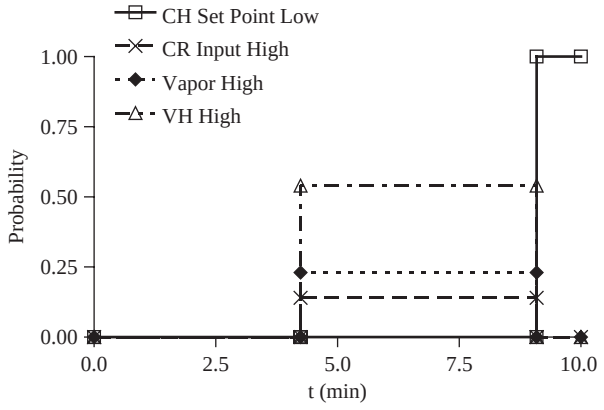

Fig. 4. Diagnostic for fault #CH Set Point Low.

Table 5
Qualitative deviations for fault #CH Output Low

t (min)	T0s	AT0
0	0	0
9.70	0	−1

50% of magnitude and 50% of development, and Fig. 5 shows the corresponding classifier output. In this case, the classifier receives only one symptom, which is not enough to identify the simulated fault. The same problem was reported by Tarifa and Scenna [9]; they employed an expert system as classifier, which also failed to identify the aforementioned fault. However, in Fig. 5, the faults reported with high probabilities are located in the plant sector corresponding to the simulated one, which is a convenient result because at least the right plant sector was identified.

The analyzed MSF desalination plant was also studied by previous works. As was mentioned above, Tarifa and Scenna [9] employed an expert system. The expert system rules were obtained automatically from a SDG (Signed Directed Graph) by using qualitative simulation. The rules were evaluated by using fuzzy logic. The selected sensors were the same. The detector block was the same one used in this work, and the ''first change'' approach was also used. The performance of that diagnostic system was comparable to the performance of the diagnostic system proposed in this work for the particular analyzed plant. However, in favor of the first diagnostic system are its capability to generate natural explanation for the obtained results, and its potential to obtain better resolution (discrimination) of faults. In favor of the diagnostic system proposed in this work are its simplicity and its capacity to give faster results. Those differences are due to the nature of the model used in each work. In fact, the SDG takes into account the sequence of the symptoms, whereas the Naive Bayes classifier does not. The complementary characteristics of both diagnostic systems suggest the convenience of combining both.

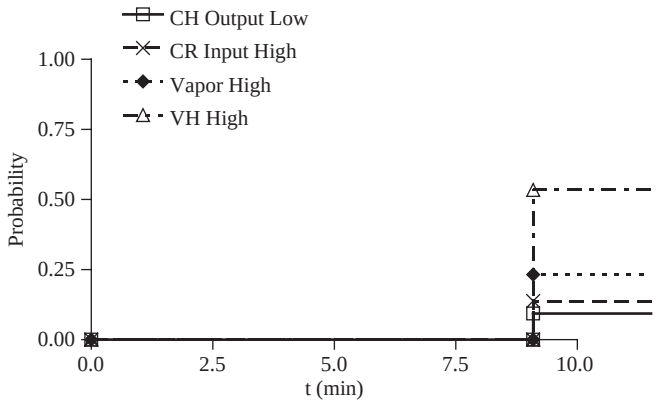

Fig. 5. Diagnostic for fault #CH Output Low.

Other work studied the same MSF desalination plant by using ANNs (Artificial Neural Networks) [10]. The selected sensors were the same. The detector block produces quantitative deviations instead of qualitative ones, and the ''first change'' approach was not used. The performance of that diagnostic system was better than the performance of the diagnostic system proposed in this work for the particular analyzed plant. However, in favor of the diagnostic system proposed in this work are still its simplicity and its capacity to give faster results. Those differences are due to the nature of the model used in each work. In fact, the ANNs take into account both the sequence and magnitude of the symptoms, whereas the Naive Bayes classifier does not. The complementary characteristics of both diagnostic systems suggest the convenience of combining both.

5. Conclusions

A diagnostic system for a MSF plant was presented. The variables to be analyzed were selected to enable an early fault detection and discrimination. Those variables were transformed in qualitative deviations by a detector block. The qualitative deviations were entered into a Naive Bayes classifier. The system performance was evaluated by using a dynamic simulator. The obtained results agree with previous works in the sense that a Naive Bayes classifier has high efficacy in spite of its simplicity. Moreover, the wrong diagnostics observed during the evaluation were caused by lack of information, not by an intrinsic limitation of the classifier. Considering that, further research is needed to generate additional information from plant data.

The proposed diagnostic system was compared with two alternatives systems. That comparison showed the convenience of combining the studied systems due to their complementary characteristics.

Acknowledgment

The authors wish to acknowledge the financial support of the Consejo Nacional de Investigaciones Científicas y Técnicas CONICET (Argentina) and Universidad Nacional de Jujuy UNJu (Argentina).

List of symbols

δX vector of qualitative or quantitative standardized deviations of process variables.

$\triangle t$ sampling period.

$\triangle Xn$	vector of normal half bands of process variables.
j	subscript that represents a process variable.
k	subscript that represents a sampling period.
p	vector of fault probabilities.
t	sampling time.
t_0	detecting time, at which the first symptom is observed.
X	vector of quantitative values of process variables.
Xn	vector or array of normal values of process variables.

References

[1] V. Venkatasubramanian, R. Rengaswamy, K. Yin and S.N. Kavuri, A review of process fault detection and diagnosis Part I: Quantitative model-based methods. *Comput. Chem. Eng*., 27 (2003a) 293–311.

[2] V. Venkatasubramanian, R. Rengaswamy and S.N. Kavuri, A review of process fault detection and diagnosis Part II: Qualitative models and search strategies. *Comput. Chem. Eng*. 27 (2003b) 313–326.

[3] V. Venkatasubramanian, R. Rengaswamy, S.N. Kavuri and K. Yin, A review of process fault detection and diagnosis Part III: Process history based methods. *Comput. Chem. Eng*., 27 (2003c) 327–346.

[4] H. Su and H. Dong, Transformer fault diagnosis based on reasoning integration of rough set and fuzzy set and Bayesian optimal classifier. WSEAS Trans. Circ. Syst., 8 (1) (2009) 136–145.

[5] D. Zhu, J. Bai and S.X. Yang, A multi-fault diagnosis method for sensor systems based on principle component analysis. Sensors, 10(1) (2010) 241–253.

[6] O. Doguc and J.E. Ramirez-Marquez, Using Bayesian approach for sensitivity analysis and fault diagnosis in complex systems. J. Integr. Design Process Sci., 13 (1) (2009) 33–48.

[7] R. Abreu, P. Zoeteweij and A.J.C. van Gemund, A new Bayesian approach to multiple intermittent fault diagnosis. 21st International Joint Conference on Artifical Intelligence, Pasadena, California, USA, 2009, pp. 653–658.

[8] J. Cheng, R. Greiner, J. Kelly, D. Bell and W. Liu, Learning Bayesian networks from data: An information-theory based approach. Artif. Intell., 137 (1–2) (2002) 43–90.

[9] E.E. Tarifa and N.J. Scenna. Fault diagnosis for a MSF using a SDG and fuzzy logic. Desalination, 152 (2002) 207–214.

[10] E.E. Tarifa, D. Humana, S. Franco, S. Martínez, A. Núñez and N. Scenna, Fault diagnosis for a MSF using neural networks. Desalination, 152 (2002) 215–222.

[11] E.E. Tarifa and N.J. Scenna, A Dynamic Simulator for MSF Plants. Desalination, 138 (2001) 349–364.

[12] J. Cheng and R. Greiner, Comparing Bayesian Network Classifiers. *15th International Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, 1999.

[13] M.J. Pazzani, Searching for dependencies in Bayesian classifiers. In Fisher D., Lenz H.J., eds., Learning from Data: Artificial Intelligence and Statistics V, Springer Verlag, 1996, pp. 239–248.

[14] H. Zhang, Exploring conditions for the optimality of Naïve Bayes. Int. J. Pattern Recog. Artif. Intell., 19 (2) (2005) 183–198.