FDI in Multivariate Process with Naive Bayesian Network in the Space of Discriminant Factors

Teodor Tiplica, Sylvain Verron, Abdessamad Kobi, Iulian Nastac

FDI in Multivariate Process with Naïve Bayesian Network in the Space of Discriminant Factors

Teodor Tiplica[*], Sylvain Verron[*], Abdessamad Kobi[*], Iulian Nastac[**]
e-mail : teodor.tiplica@istia.univ-angers.fr

Abstract

The Naïve Bayesian Network (NBN) classifier is an optimal classifier(in the sense of minimal classification error rate) in the case of independent descriptors or variables. The presence of dependencies between variables generally reduce his efficiency. In this article, we are proposing a new classification method named Naïve Bayesian Network in the Space of Discriminants Factors (NBNSDF) which is based on the use of the NBN in the space of discriminants factors issue from a discriminant analysis. The discriminants factors are not correlated letting very efficient the use of the NBN. We found on simulated data that the NBNSDF method better detects and isolates faults in multivariate processes than the NBN in the case of strongly correlated variables.

1. Introduction

Nowadays, in all the manufacturing processes we have to deal with big amounts of data. Therefore, it becomes difficult to extract the useful information about the state of the process, in order to take the right decision when problems are encountered. To do that, we are using knowledge of all kinds. In this context, the fault detection and isolation (FDI) techniques and the classification techniques take an important place. The classification requires the construction of a classifier (a function assigning a class identifier to the observations which are described by a certain number of descriptors).

It was theoretically demonstrated [16, 17] that, according to the criterion of classification error rate, there is no universal algorithm always exceeding the others. In practice, other criterions can be retained, in order to compare classification methods, like: the effectiveness, the comprehensibility, the training time, etc, see [8]. Given these comparison criterions, one can define an utility function to try to discriminate them according to the specificity of the problem which is treated [3]. A very interesting comparative study of some classification algorithms, frequently used by the community of the artificial intelligence, was carried out by Inza et al. [6].

In the context of the supervised classification, the Naïve Bayesian Network (NBN) classifier was the subject of a particular attention. Its performances were analyzed and compared with those of some very well known statistical methods like the hierarchical classification, the K nearest neighbors, the C4.5, the decision trees, see [9, 12].

A great advantage of the NBN classifier, except its simplicity, is its low sensitivity to the noise (not-informative variables). If the independence assumption of the descriptive variables is valid, it was shown [18], that the NBN classifier is optimal, which means that he has the lowest classification error rate. It should nevertheless be mentioned here that, in almost all of the practical situations, the independence assumption of the descriptive variables is very unrealistic.

The presence of correlations between the descriptors can reduce the efficiency of the NBN, see [10]. However, even in the case of correlated variables, we can affirm that the NBN classifier gives very good results compared to other more sophisticated classifiers, see [4].

We are proposing in this article a new classification method whose performances for FDI are higher than those of the NBN classifier. The principle of the method that we are proposing differs substantially from those of the other methods published in the literature. Just like the NBN classifier in the case of the not-correlated variables, our classifier wants to be optimal (in terms of classification error rate) but in the case of the correlated descriptive variables.

[*] LASQUO Laboratory of ISTIA, University of Angers (62, Avenue Notre Dame du Lac, 49000 Angers, France)
[**] CAQ Laboratory of Electronic and Telecommunications Faculty of Bucarest (Bd. Iuliu Maniu nr. 1-3; Bucharest; Romania)

Our paper is structured in the following manner: in the second paragraph we are briefly presenting a non-exhaustive set of classifiers derived from the NBN classifier model and proposed in the literature to treat the case of the correlated variables; the third paragraph introduces some guiding principles of the discriminant analysis technique; in the fourth paragraph we are presenting the principle of the classification method that we are proposing; in the fifth paragraph we are illustrating the efficiency of our classifier in the frame of the FDI in multivariate processes; finally in the sixth paragraph we are concluding and we are mentioning some prospects for our work.

2. Bayesian Network Classifiers

Given the increasing interest caused by the NBN classifier in the field of the artificial intelligence, many researches have been carried out in the last years in order to improve its performances (expressed as classification accuracy) in the case of correlated descriptive variables. Madden [12] carries out a very good comparative analysis of these techniques.

A first solution may be the utilization of the Bayesian Networks (BN) for classification tasks. It has been shown (see [4]) that, if the number of descriptors is reduced, this approach has better performances compared to the NBN classifier. Unfortunately, the efficiency of the BN decreases when the number of the descriptors increases. Thereby, certain solutions were proposed in order to increase the efficiency of the BN when faced to a great number of descriptors. They can be divided in two great families : dimensional reduction and partial correlation.

2.1 Dimensional reduction

Dimensional reduction of the descriptors space keeps only the most representatives and the most informative variables for the classification. According to this criterion, variables are divided in three categories: strongly representatives (variables leading to the increase of the classification error rate if ignored); weakly representatives (variables contributing sometimes to the reduction of the classification error rate) and not-representatives for the classification task (variables which are neither strongly representatives or weakly representatives), see [7]. To reduce the dimension of the descriptors space we need a metric (for example: the classification error rate) allowing to evaluate and to compare the different groups of descriptive variables. We also need a searching algorithm in the space of the descriptive variables, in order to identify the most adequate group of descriptors, according to the adopted criterion. In this context, several heuristic searching algorithms (for example: the backward elimination algorithm or the forward selection algorithm) were proposed. The Selective Bayesian Classifier proposed by Langley [10] is particularly interesting in the case where the descriptive variables are strongly correlated. Langley's approach consists in extracting a subset of variables from the whole set of variables in order to eliminate the important correlations.

2.2 Partial correlation

Partial correlation allows, under constraint, the existence of certain correlations between the descriptive variables. The Tree Augmented Naïve Bayes (TAN) classifier, proposed by Friedman [4], was derived from the NBN classifier. A certain number of arcs are added to the NBN in order to take into account the correlations between the descriptive variables. This generalization leads to an increasing complexity of calculation because, it is necessary to search in the space of all possible networks the one which fits the better to the structure of data. Nevertheless, by imposing some constraints to the network structure (like for example: the class node does not have parents; each descriptor node has like parents the class node and at least one more another descriptor node) the computational time can be rather reasonable. The network structure is found by using a modified MDL (Minimum Description Length) algorithm which takes into account the constraints mentioned before. A generalization of the TAN classifier is obtained by considering that the correlation between the descriptors can change from one class to another. In this case, one can build a TAN classifier for each class. This set of local networks is named Bayesian Multinet. It was shown, see [4], that the TAN and the Bayesian Multinet classifiers have similar and often better performances (in term of classification error rate) than the NBN classifier, especially for the case where the number of classes is important.

3. Discriminant Factors computation

The discriminant analysis is a family of statistical techniques whose objective is to assign individuals (characterized by a certain number of numeric or nominal variables) to preexistent classes, see [11]. The data consist of n observations divided into q classes and described by p predictors. One can distinguish two

major aspects in the discriminant analysis: **descriptional** (identify the linear combinations of the descriptive variables allowing to best separate the existing classes) and **decisional** (decide in which class it is better to assign a new individual for whom one knows the values of the predictors). The n observations (or individuals) x_i constitute a group of dots Y in the p-dimensional space R^P divided into q subgroups Y_1, Y_2,...,Y_q having their gravity centers g_1, g_2,...,g_k.
The theorem of Huygens enables us to write the following equation:

$$T = B + W \quad (1)$$

where: $T = \frac{1}{n} \sum_{k=1}^{q} \sum_{x_i \in Y_k} x_i^T x_i$ is the total variance covariance matrix;

$W = \frac{1}{n} \sum_{k=1}^{q} \sum_{x_i \in Y_k} n_k (x_i - g_k)^T (x_i - g_k)$ is the within classes covariance matrix, and $B = \sum_{k=1}^{q} \frac{n_k}{n} g_k^T g_k$ is the between classes covariance matrix.

The projection of the equation 1 on a linear form u defined on R^P is given by the following equation:

$$uTu^T = uWu^T + uBu^T \quad (2)$$

The first objective of the discriminant analysis technique is to find the factorial axis u which separates as well as possible the existing classes. The separation of the classes is better if the classes are distant (large between class variance) and the observations of the same class are close (small within class variance). Consequently, the first discriminant factorial axis u_1 will be the element which maximizes the following ratio:

$$\max_{(u)} \left(\frac{uBu^T}{uTu^T} \right) \quad (3)$$

It can be demonstrated that the first discriminant factorial axis u_1 is the eigenvector of the matrix $T^{-1}B$ corresponding to the greatest eigenvalue λ_1. In the same way, one defines the second discriminant factorial axis u_2 like the second eigenvector of the matrix $T^{-1}B$ and orthogonal with u_1. The second linear form u_2 constitutes the best discriminating factor independent of u_1.

In discrimination problems, the maximum number of discriminant factorial axis is equal to the number of groups minus one (q-1), or the number of variables in the analysis (p), whichever is smaller. Consequently, under the assumption that $p > q$, the individuals x_i generate the R^P space, while the discriminant axis generate the lower dimensional space R^{q-1}. In this way, one can says that sometimes the transformation imposed by the discriminant analysis contributes also to the dimensional reduction of the descriptors space.

4. New Classification Method

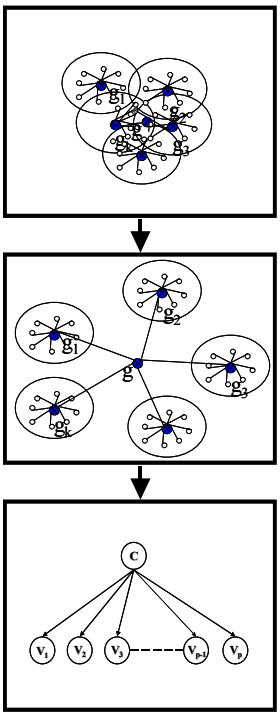
The classification method that we are proposing in this article, named Naive Bayesian Network in the Space of Discriminants Factors (NBNSDF), has a completely different orientation compared to the methods previously quoted (see the second paragraph). Our approach wants to be optimal (lowest classification error rate) in the case of correlated variables.

The problem is then summarized in the following way: we have to find a solution to transform the correlated variables in completely independent variables in order to carry out thereafter the classification using the NBN classifier. For that, we have to transform the space of the descriptive variables to lead to a new space where the new descriptive variables are not correlated. This transformation is done by the discriminant analysis when the discriminating factors are calculated.

Then, by projecting the set of observations on the discriminating factorial axes we obtain the new independent descriptors used for the classification. This projection generally leads also to a dimensional reduction of the descriptor's space because, as we specified in the third paragraph, the number of discriminating factorial axes is equal to q-1 (where q is the number of classes), and very often $q < p$. Thereafter, a NBN is trained to recognize the classes in the new space R^{q-1}. Each node of this network corresponds to a discriminant factorial axis and the values which it can take represent the projections of the former variables (forming the space R^p) on this axis (see figure 1).

By using the NBNSDF classifier we don't need to reduce the number of descriptive variables and thus to lose information which could be useful for the fault detection and isolation. Also, we don't need to take into account only certain correlations of the descriptive variables and thus to simplify too much the analyzed

model, which would lead sometimes to an important gap between the model and the real process.



Step 1: The set of observations in the original space defined by the p descriptors.

Step 2: The same set of observations in the transformed space defined by the q – 1 discriminating factorial axis.

Step 3: Performing the classification by using the NBN classifier in the q-1 dimensional space defined by the discriminating factorial axis.

Figure 1. Steps of the NBNSDF classification methodology

5. The performances of the NBNSDF Classifier

We compared the performances of the NBNSDF classifier with those of the NBN and ANN (Artificial Neural Network) classifiers in the case of the FDI in multivariate processes. We choused to compare the performances of the NBNSDF method with those of the ANN because the ANNs have some interesting classification properties : ANNs are free of any distributional assumptions, can deal with inter-correlated data, and they provide a mapping function from the input to the outputs without any a priori knowledge about the function form since they are universal approximators [5]. A comprehensive study on ANN for failure prediction is [15] where the author investigates fifteen related papers for a number of characteristics: what data was used, what types of ANN models, what software, what kind of network architecture, etc.

In this work, four scenarios of simulation which cover all the possible correlation structures of variables were used : in the first case the variables are not correlated; in the second case the variables are correlated inside each class of fault but are not correlated globally; in the third case the variables are correlated globally but are not correlated inside each class of fault and finally in the fourth case the variables are correlated globally and also inside each class of fault.

The 50 simulations per scenario which we carried out in this study relate to four descriptive variables (or process parameters) normally distributed with zero mean and standard deviation equal to one but, obviously, our method is not limited by the number of the variables to use in the analysis. For each simulation, we used two groups of observations: in the first group (used for learning the network) we simulated various faults and in the second group (used for testing the network) we reproduced the same faults and noted the classification error rate of the classifier.

The training set of data which was used here comprises 30 observations for each type of fault and 30 observations for the normal operating conditions of the process.

Seven faults, representing positive and negative amplitude shifts of the mean of each variable, were simulated (see table 1 and 2) for each scenario. The A parameter represents the amplitude of the mean shift and can take the following values: 2; 2.5 and 3.

x_1	x_2	x_3	x_4	Fault code
0	0	0	0	Normal
0	0	A	A	Fault 1
0	0	A	-A	Fault 2
A	A	0	0	Fault 3
A	-A	0	0	Fault 4
A	-A	A	-A	Fault 5
-A	-A	-A	-A	Fault 6

Table 1. The simulated fault matrix for not-correlated variables

x_1	x_2	x_3	x_4	Fault code
0	0	0	0	Normal
0	0	A	A	Fault 1
0	0	-A	-A	Fault 2
A	A	0	0	Fault 3
-A	-A	0	0	Fault 4
A	A	A	A	Fault 5
-A	-A	-A	-A	Fault 6

Table 2. The simulated fault matrix for correlated variables

The following variance-covariance matrices were used for each simulated scenario :

Case 1: $\Sigma_{global} = \Sigma_{classe} = 0.8 \times I_{4\times4} + 0.2 \times 1_{4\times4}$

Case 2: $\Sigma_{global} = 0.8 \times I_{4\times4} + 0.2 \times 1_{4\times4}$

$\Sigma_{classe} = 0.2 \times I_{4\times4} + 0.8 \times 1_{4\times4}$

Case 3: $\Sigma_{global} = 0.2 \times I_{4\times4} + 0.8 \times 1_{4\times4}$

$\Sigma_{local} = 0.8 \times I_{4\times4} + 0.2 \times 1_{4\times4}$

Case 4: $\Sigma_{global} = \Sigma_{classe} = 0.2 \times I_{4\times4} + 0.8 \times 1_{4\times4}$

where: Σ_{global} – represents de global variance-covariance matrix; Σ_{classe} – represents de variance-covariance matrix of each simulated class of fault; $I_{4\times4}$ – represents the 4 × 4 identity matrix and $1_{4\times4}$ – represents the 4 × 4 ones matrix.

The construction of the conditional probabilities tables for the BN generally imposes the discretization of the continuous descriptors. The discretization results in cutting the continuous descriptor in a certain number of intervals and allocating a single value (which represents the interval) to any value of the descriptor belonging to that interval. In this study, we used the discretization in 10 equal width intervals for the continuous variables. Even if this discretization method is not the most powerful for the classification task, see [2, 14], the equal width discretization method gives very satisfactory results for the NBN classifier [18].

The results obtained following 50 simulations are summarized in table 3. Each case has been tested with different algorithms : the Bayes classifier (NBN), the Bayes classifier in the space of discriminant factors (NBNSDF) and an artificial neural network (ANN).

We used feed forward ANNs with two hidden layers in order to achieve a good classification function, based on our preliminary research, where we have obtained better results in case of two hidden layers than in case of one hidden layer, however maintaining a similar ratio (approx. 2/1, and greater than 1/1) between the number of the training samples and the total number of the weights.

The Scaled Conjugate Gradient (SCG) algorithm [13] was used for training the network. In order to avoid the overfitting phenomenon, we applied the early stopping method (validation stop) during the training process. As the splitting criterion, we randomly choose approximately 85% of the data for training set, and the rest for validation. Furthermore, we imposed the supplementary condition:

$$E_{val} \leq \frac{6}{5} \cdot E_{tr} \quad (4)$$

to avoid a large difference between the error of the training set and the error of the validation set. In this way, the overfitting phenomenon on the test set will be considerably reduced. In our approach the validation set acts at the same time as a kind of test set, even though there is a real and separate test set for different input-output pairs.

In table 3 we represented the mean accuracy of the 50 simulations and the standard deviation in parenthesis. Bold characters represent the best accuracy for the simulation at a significance level of 1%.

Case 1

A	NBN	NBNSDF	ANN
2	78.5 (2.67)	77.05 (2.82)	**80.45 (3.15)**
2.5	87.44 (2.52)	87.23 (2.53)	**89.46 (2.71)**
3	92.82 (1.92)	91.72 (1.86)	**93.81 (2.35)**

Case 2

A	NBN	NBNSDF	ANN
2	78.02 (3.39)	**93.51 (1.78)**	**93.70 (2.09)**
2.5	84.80 (3.07)	**95.66 (1.36)**	**95.96 (1.40)**
3	88.30 (2.45)	**96.86 (1.51)**	**96.93 (1.19)**

Case 3

A	NBN	NBNSDF	ANN
2	73.71 (3.13)	74.59 (2.80)	**79.5 (2.79)**
2.5	84.51 (2.45)	83.94 (2.81)	**88.01 (2.32)**
3	91.45 (2.01)	89.61 (2.18)	**92.81 (2.31)**

Case 4

A	NBN	NBNSDF	ANN
2	64.42 (3.37)	**78.41 (3.2)**	**77.89 (3.09)**
2.5	78.1 (3.09)	**86.31 (2.66)**	**85.01 (2.91)**
3	85.88 (2.57)	**90.15 (2.17)**	**90.80 (2.27)**

Table 3. Classification accuracy (comparative results)

We can see that the accuracy of each classifier increases when the magnitude of the shift increases. That is not surprising because classes are as more separated of each others in the multidimensional space as the magnitude of the shift increases. We can also note that NBN and NBNSDF have quite similar results in the cases of not correlated variables inside each class of faults (case 1 and case 3). This result is due to the fact that the discriminant analysis do not improve the classes separation when dealing with completely not correlated variables. In these two cases, ANN have the best accuracy. It is interesting to see that the global

correlation between variables does not improve significantly the classification accuracy of the NBNSDF classification method (case 3 versus 1 and case 4 versus 2). It seems even that the correlation between the variables induces worst results in term of accuracy rate. But we must not forget that the faults matrices are not the same for the globally correlated and not-correlated variables (faults 2, 4 and 5 are not the same in these scenarios – see tables 1 and 2) and this could be at the origin of this difference (some faults are probably more easy to classify than others). Finally, in the cases of correlated variables inside each class (cases 2 and 4), NBNSDF have a better accuracy than the NBN. His accuracy is similar to that of the ANN. But we should mention here that the NBNSDF classifier have easiest computational simplicity compared to ANN which learning is long and not evident.

6. Conclusion

In this article we are proposing a new classification method named NBNSDF which is based on the use of the NBN classifier in the space of the discriminant factors resulting from a discriminant analysis. We compared the performances of the NBNSDF classifier with those of the NBN and ANN classifiers on some simulated FDI scenarios. We could note that, the performances of the NBNSDF classifier are quite the same as those of the NBN classifier in the cases of not correlated variables inside each class of faults. But, we improved in a substantial way the performances of the NBN classifier in the case where variables are correlated. This was, as mentioned at the beginning of this paper, one of the main objective of our work. Even if, in the case of correlated variables, the NBNSDF and the ANN classifiers have the same performances, the NBNSDF classifier has the advantage to be more easiest to compute, more faster an more simple than the ANN classifier.

Future work has to be done to analyze how the classification accuracy is influenced by the number of observations in each class of fault, by the number of the descriptive variables and by the number of fault classes. Also, work is still remaining to analyze which is the influence of the correlation coefficient between the variables on the classification accuracy.

References

[1] B. R. Bakshi. “Multiscale pca with application to multivariate statistical process monitoring”. *AIChE Journal*, 44(7):1596. 1610, 1998.

[2] S. D. Bay. “Multivariate discretization of continuous variables for set mining”. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. “From data mining to knowledge discovery in databases”. *AI Magazine*, 17(3):37.53, 1996.

[4] N. Friedman, D. Geiger, and M. Goldszmidt. “Bayesian network classifiers”. *Machine Learning*, 29(2-3):131.163, 1997.

[5] Hornik, K., Stinchcombe, M. and White H. “Multilayer feedforward networks are universal approximators”, *Neural Networks*, 2:359.366, 1989.

[6] I. Inza, P. Larranaga, B. Sierra, R. Etxeberria, J. Lozano, and J. Pena. “Representing the behaviour of supervised classification learning algorithms by bayesian networks”. *Pattern Recognition Letters*, 20(11-13):1201.1209, 1999.

[7] G. H. John, R. Kohavi, and K. Pfleger. “Irrelevant features and the subset selection problem”. In *International Conference on Machine Learning*, pages 121.129, 1994.

[8] R. Kohavi, D. Sommerfield, and J. Dougherty. “Data mining using mlc++. a machine learning library in c++”. *International Journal on Artificial Intelligence Tools*, 6(4):234.245, 1997.

[9] P. Langley, W. Iba, and K. Thompson. “An analysis of bayesian classifiers”. In *National Conference on Artificial Intelligence*, 1992.

[10] P. Langley and S. Sage. “Induction of selective bayesian classifiers”. *In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 1994.

[11] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. DUNOD, 2000.

[12] M. G. Madden. “ The performance of bayesian network classifiers constructed using different techniques”. In *Proceedings of European Conference on Machine Learning, Workshop on Probabilistic Graphical Models for Classification*, September 2003.

[13] M. Moller. “A scaled conjugate gradient algorithm for fast supervised learning”. *Neural Networks*, 6:525.533, 1993.

[14] S. Monti and G. Cooper. “A multivariate discretization method for learning bayesian networks from mixed data”. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, 1998.

[15] O’Leary, D.E. (1998), “Using Neural Networks to Predict Corporate Failure”, *International Journal of Intelligent Systems in Accounting, Finance & Management* 7:187.197, 1998.

[16] C. Schaffer. “A conservation law for generalisation performance”. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 259.265, 1994.

[17] D. Wolpert. “The relationship between pac, the statistical physics framework, the bayesian framework and the vc framework.” 1994.

[18] Y. Yang and G. I. Webb. “A comparative study of discretization methods for naive-bayes classifiers”. In *Proceedings of PKAW 2002*, 2002