Chapter 2

Bayesian Approach in Medicine and Health Management

Emanuela Barbini, Pietro Manzi and Paolo Barbini

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/52402

1. Introduction

Statistical decision theory deals with scenarios where decisions have to be made under a state of uncertainty, and its goal is to provide a rational framework for dealing with such situations. These scenarios are typical in most of the problems of medical decision-making.

The Bayesian procedure is a particular way of formulating and dealing with these type of problems. It has great promise in putting health-related decision making on a more rational basis, thus making the assumptions more obvious, and making the decisions easier to explain and defend [1]. This approach can be used to support the decision-making process in many application fields, as, for example, diagnosis and prognosis [2], risk assessment [3] and health technology assessment [4]. A wide-ranging collection of applications of Bayesian statistics in the biomedical field can be found in thematic books [5-7].

Bayes's theorem appeared in a posthumous publication by Thomas Bayes [8]. It gives a simple and uncontroversial result in probability theory, but its practical application has been the subject of considerable controversy for more than two centuries and, only recently, a more balanced and pragmatic perspective has emerged. In summary, the Bayesian approach offers a method of formalizing a priori beliefs and of combining them with the available observations, with the goal of allowing a rational derivation of optimal decision criteria.

This chapter is an introduction to the modern world of Bayes procedures for professionals with a minimal background in biostatistics. In particular, we analyze the main theoretical aspects of the Bayesian approach and we review current thinking on the value of this approach as a decision-support system in medicine and health management. It is hoped that this discussion stimulates reflection on the subject and elicits new ideas and inspiration for people involved in problems of public health.

2. The Bayes decision rule

Statistical decision-making can be seen as a process of inferring, from past observations, predictions that then can be used to perform an action. A typical problem is to determine to which class a given sample belongs. The estimation of the class membership can provide a key support to decision making.

The first step of the process is the formalization of the underlying unknown reality (class membership). This is done by considering that this unknown reality can be represented by an entity (ω) taking values on a state space (Ω). Often, ω is a single unknown numerical quantity. In other problems, it may be an ordered set of numerical parameters (a vector) or the elements of Ω may not be of numerical nature. In the present chapter, we assume that ω is a single quantity.

An observation and/or measurement process allow us to obtain a set of numbers which make up the observation vector ($\mathbf{x}$). This observation vector is assumed to be a random vector whose conditional density function depends on ω. In formal probabilistic terms, this dependence is expressed by the class-conditional probability function $p(\mathbf{x}|\omega)$. The observations $\mathbf{x}$ are also called features and the feature vector is the input to the decision rule by which the sample is assigned to one of the given class.

The Bayesian approach to decision theory brings into play another element: *a priori* knowledge which concerns ω, in the form of a probability function $P(\omega)$. This probability is usually referred to as the prior.

When the prior probabilities and the class-conditional probability functions are known it is possible to derive a formal decision rule, which allows us to decide to which class a given sample ($\mathbf{x}$) belongs. For simplicity of discussion we will treat the two-class problem. The generalization to more than two classes is immediate.

Let $\mathbf{x}$ be the observation vector and let be ω_1 and ω_2 the two classes to which $\mathbf{x}$ may belong. A decision rule based on probabilities can be written as follows [9]

$$\begin{aligned} P(\omega_1 \mid \mathbf{x}) \geq P(\omega_2 \mid \mathbf{x}) \quad &\Rightarrow \quad \mathbf{x} \in \omega_1 \\ P(\omega_1 \mid \mathbf{x}) < P(\omega_2 \mid \mathbf{x}) \quad &\Rightarrow \quad \mathbf{x} \in \omega_2 \end{aligned} \tag{1}$$

where $P(\omega_i|\mathbf{x})$ is the conditional probability of ω_i given $\mathbf{x}$ (i=1,2), that is the probability of ω_i if $\mathbf{x}$ has occurred. This conditional probability is also known as the *a posteriori* probability of ω_i given $\mathbf{x}$.

Bayes's theorem allows us to write the *a posteriori* probabilities of ω_1 and ω_2 as a function of the *a priori* probabilities ($P(\omega_1)$ and $P(\omega_2)$) and the class-conditional probability functions ($p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$), giving

http://dx.doi.org/10.5772/52402

$$P(\omega_i \mid \mathbf{x}) = \frac{P(\omega_i) \times p(\mathbf{x} \mid \omega_i)}{p(\mathbf{x})} \qquad i = 1, 2 \tag{2}$$

where p(**x**) is the probability density function of **x**.

In the case of two mutually exclusive classes, p(**x**) can be expressed as

$$p(\mathbf{x}) = p(\mathbf{x} \cap \omega_1) + p(\mathbf{x} \cap \omega_2) = P(\omega_1) \times p(\mathbf{x} \mid \omega_1) + P(\omega_2) \times p(\mathbf{x} \mid \omega_2) \tag{3}$$

Therefore, the posterior probabilities in equation 2, which determine the decision rule in equation 1, can be calculated on the basis of the prior and class-conditional probabilities. This confirms that the Bayesian approach, combining *a priori* beliefs with available observations, updates the knowledge in the light of experience.

Combining equation 1 with equation 2, the previous decision rule based on probabilities can be rewritten as follows

$$\begin{aligned} P(\omega_1) \times p(\mathbf{x} \mid \omega_1) \geq P(\omega_2) \times p(\mathbf{x} \mid \omega_2) \quad &\Rightarrow \quad \mathbf{x} \in \omega_1 \\ P(\omega_1) \times p(\mathbf{x} \mid \omega_1) < P(\omega_2) \times p(\mathbf{x} \mid \omega_2) \quad &\Rightarrow \quad \mathbf{x} \in \omega_2 \end{aligned} \tag{4}$$

or, equivalently,

$$\begin{aligned} l(\mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} \geq \frac{P(\omega_2)}{P(\omega_1)} \quad &\Rightarrow \quad \mathbf{x} \in \omega_1 \\ l(\mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} < \frac{P(\omega_2)}{P(\omega_1)} \quad &\Rightarrow \quad \mathbf{x} \in \omega_2 \end{aligned} \tag{5}$$

The term l(**x**) is called the likelihood ratio and the ratio $P(\omega_2)/P(\omega_1)$ is the threshold value of the likelihood ratio for the decision.

Equation 5 corresponds to the Bayes's decision rule for minimum error and defines the decision boundary between the two classes. In particular, when $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ are Gaussian with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the decision boundary is linear. Analogously, when $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ are Gaussian with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and covariance matrices $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, the decision boundary is quadratic. Finally, when $p(\mathbf{x}|\omega_1)$ and/or $p(\mathbf{x}|\omega_2)$ are not Gaussian, the shape of the decision boundary cannot be defined a priori. Figures 1 and 2 show two-dimensional examples for Gaussian class-conditional probability functions when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, respectively.

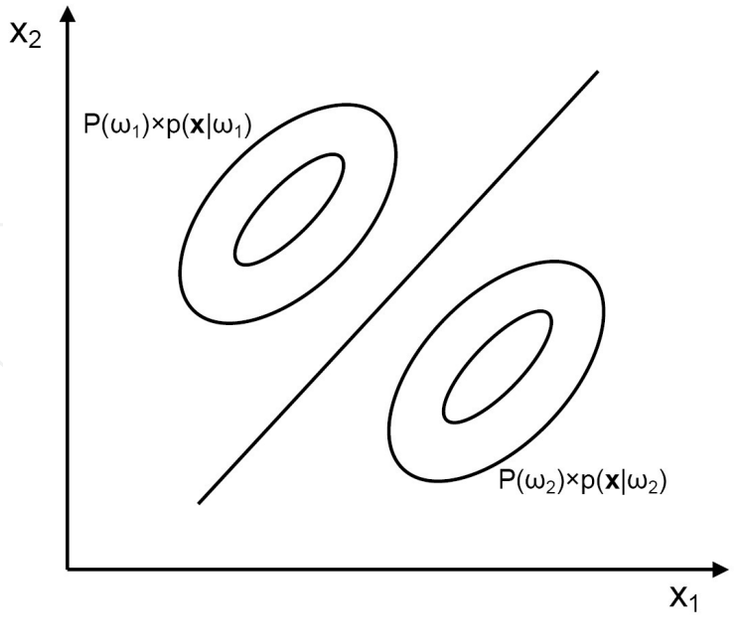

Figure 1. Decision boundary for Gaussian class-conditional probability functions when $\Sigma_1 = \Sigma_2$. Ellipses depict contour lines.

The decision rule of equation 1 implies that we should choose the class whose posterior probability is largest. Bearing in mind that the sum of $P(\omega_1|\mathbf{x})$ and $P(\omega_2|\mathbf{x})$ is equal to 1, this means setting the posterior class-conditional probability threshold (P_t) equal to 0.5, that is, $\mathbf{x}$ is assigned to class ω_i if $P(\omega_i|\mathbf{x})$ is greater than 0.5. However, the decision rule may be formulated using somewhat different reasoning. Often, in medical applications, the decision rule must account for the cost of a wrong decision. In this case, a cost can be assigned to each correct and wrong decision and P_t is chosen to obtain the minimum risk decision rule. From a purely mathematical point of view, the selection of costs is equivalent to a change in prior probabilities.

The discrimination capacity of a classification model assesses model performance in assigning correctly a sample to a class. Many criteria exist for evaluating discrimination capacity. A common way for the two-class problem is to evaluate sensitivity (SE) and specificity (SP). SE can be interpreted as the fraction of cases from ω_1 (for example, diseased patients or high-risk subjects) which are correctly classified, while SP gives the fraction of correctly classified cases from ω_2 [10]. Generally, SE and SP depend on the chosen probability threshold P_t to which the posterior probability is compared.

A receiver operating characteristic (ROC) curve is a graphic representation of the relationship between the true-positive fraction (TPF=SE) and false-positive fraction (FPF=1-SP) obtained for all possible choices of P_t.

http://dx.doi.org/10.5772/52402

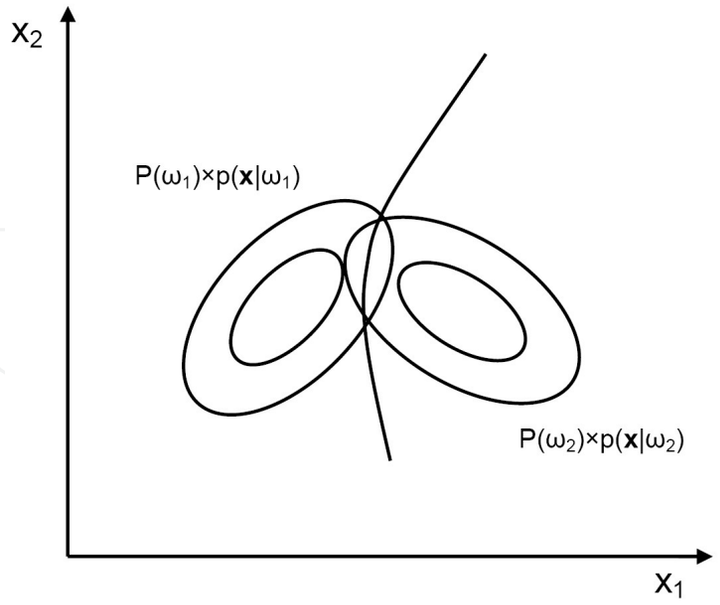

Figure 2. Decision boundary for Gaussian class-conditional probability functions when $\Sigma_1 \neq \Sigma_2$. Ellipses depict contour lines.

In medical applications the area under the ROC curve (AUC) is the most commonly used global index of discrimination capacity, although alternative indices can be used [11]. The AUC can be interpreted as the chance that a case randomly selected from ω_1 will be correctly distinguished from a randomly selected case from ω_2. An AUC value equal to 1 implies perfect forecasts, while an AUC of 0.5 reflects random forecasts.

3. Estimation of class-conditional probability functions

The Bayesian decision rule provides an optimal solution to the classification problem when the underlying probabilistic structure is known. Actually, this occurs very rarely. When the probabilistic structure is not known, the most common approach to solve the problem is to estimate it from a set of available experimental data. This set of samples is called training set. Thus the class-conditional probability functions are estimated from the training set and the Bayes's rule is implemented using the estimated functions. It should however be emphasized that the Bayesian decision rule obtained in this way does not maintain its optimality characteristics.

If we can assume that the class-conditional probability functions have a well-defined structure which can be characterized by a finite number of parameters, we can derive the Bayesian decision rule using the parameter estimates. In this case the problem is the estimation of the set of parameters and this corresponds to a parametric approach. Unfortunately, in some cases we cannot assume a parametric form for the class-conditional probability functions and in order to apply the Bayesian decision rule, we have to somehow estimate these func-

tions, using an unstructured (nonparametric) approach. The nonparametric approach is complex and its description is not an objective of our discussion. The reader interested in this topic can find an exhaustive treatment of the problem in the specific literature [12].

Generalization is an issue of crucial importance for decision models which were designed on a training set. It is defined as the capacity of the model to maintain the same performance on data not used for training, but belonging to the same population. It is therefore estimated by testing model performance on a different set of available experimental data (testing data). The model generalizes well when the decision errors in testing and training data sets do not differ significantly. From this point of view, the optimal model is the simplest possible model designed on training data which shows the highest possible performance on any other equally representative set of testing data. Excessively complex models tend to overfit, which means they show an error on the training data significantly lower than on the testing data. Overfit is a sort of data storage precluding the learning of decision rules. It must be avoided since it causes loss of generalization.

Later in this chapter we will focus on various Bayesian models, frequently utilised in medicine and in health management, using the parametric approach to define the decision rule. After a brief description of each model, we will show an applicative example bringing out the main strengths and weaknesses of the various models.

4. The Bayes linear classifier

In many cases the class-conditional probability functions are assumed to be Gaussian, that is

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\Pi)^{q/2}} \; |\boldsymbol{\Sigma}_i|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \; \boldsymbol{\Sigma}_i^{-1} \; (\mathbf{x}-\boldsymbol{\mu}_i)\right\} \tag{6}$$

where $\boldsymbol{\mu}_i$ is the mean of class ω_i, $\boldsymbol{\Sigma}_i$ the covariance matrix, q the dimension of the observation vector (i.e. the number of features used). $|\boldsymbol{\Sigma}_i|$ indicates the determinant of $\boldsymbol{\Sigma}_i$ and superscript T indicates matrix transposition.

In this condition the problem to be solved is the estimation of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. Therefore, the number of parameters to be estimated to identify $p(\mathbf{x}|\omega_i)$ is equal to q(q+3)/2, which corresponds to q parameters of the mean vector and q(q+1)/2 parameters of the covariance matrix. The greater the number q of features, the greater the number of parameters to be estimated to define the normal class-conditional probability function. Of course, given a set of training data, the accuracy of model parameter estimates rapidly worsens as q increases, leading to a significant loss in generalization capacity.

A simplifying hypothesis that is often made is to assume that the two classes have identical covariance matrices (homoscedasticity), which, as said above, leads to obtain a linear decision boundary. In fact, in this case the decision rule in equation 5 can be rewritten as

http://dx.doi.org/10.5772/52402

$$(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\frac{1}{2}(\boldsymbol{\mu}_1^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2)\le\ln\frac{\mathrm{P}(\omega_1)}{\mathrm{P}(\omega_2)}\quad\Rightarrow\quad\mathbf{x}\in\omega_1$$
$$(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\frac{1}{2}(\boldsymbol{\mu}_1^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2)>\ln\frac{\mathrm{P}(\omega_1)}{\mathrm{P}(\omega_2)}\quad\Rightarrow\quad\mathbf{x}\in\omega_2 \tag{7}$$

which identifies the following linear boundary

$$(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{T}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\frac{1}{2}(\boldsymbol{\mu}_1^{T}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^{T}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2)=\ln\frac{P(\omega_1)}{P(\omega_2)} \tag{8}$$

where $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$.

Under this hypothesis, the total number of parameters to be estimated to define the decision rule for a two-class problem becomes q(q+5)/2.

The Bayes linear classifier is a simple approach, but, in the Bayes sense, it is optimal only for normal distributions with equal covariance matrices of the classes. However, in many cases, the simplicity and robustness of the linear model compensate the loss of performance occasioned by non-normality or non-homoscedasticity.

This type of model has been used for estimating morbidity probability after heart surgery [13]. The classifier was developed and tested analysing data acquired in a set of 1090 patients who underwent coronary artery bypass grafting and were admitted to the intensive care unit (ICU) of the University Hospital of Siena (Italy) over a period of three years (2002–2004). A collection of 78 preoperative, intraoperative and postoperative variables were considered as likely risk predictors, that could be associated with morbidity development in the ICU. A dichotomous (binary) variable was chosen as ICU outcome (morbidity). Data was ranked chronologically on the basis of patient hospital discharge and organized in a database. The database was divided into two sets of equal size (545 cases each): a training set consisting of patients in odd positions in the original ranked database and a testing set consisting of the other patients, that is, those in even positions in the original database. To ensure that alternate allocation of cases did not introduce systematic sampling errors, training and testing data was compared using statistical significance tests. Before developing the formal decision-model, a feature selection was performed, which reduced the variables from 78 to 8, thus the number of parameters estimated from the training set to identify the two class-conditional probability functions was equal to 52.

Discrimination capacity was quantified by AUC estimated on bootstrap data. For testing data, the 95% confidence interval (CI) of AUC was also calculated. Generalization was evaluated as the percentage difference in AUC between training and testing data. The AUC values were equal to 0.815 and 0.778 for training and testing sets, respectively, while CI for testing data was equal to [0.722-0.831]. The percentage difference in AUC between training and testing data was 4.5%.

5. The Bayes quadratic classifier

When the class-conditional probability functions are assumed to be Gaussian, but the two classes have not identical covariance matrices, a Bayes quadratic classifier should be employed. In this case, the decision rule in equation 5 can be expressed as a quadratic function of the observation vector **x** as

$$\begin{aligned} &\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)+\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\le\ln\frac{P(\omega_1)}{P(\omega_2)} \quad \Rightarrow \quad \mathbf{x}\in\omega_1 \\ &\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)+\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}>\ln\frac{P(\omega_1)}{P(\omega_2)} \quad \Rightarrow \quad \mathbf{x}\in\omega_2 \end{aligned} \tag{9}$$

which identifies a non linear (quadratic) boundary. The Bayes quadratic classifier in equation 9 requires the estimation of q(q+3) parameters from the training set.

Also this type of model has been used for estimating the probability of morbidity after heart surgery in the same set of 1090 patients who underwent coronary artery bypass in the University Hospital of Siena over the period 2002–2004 [13]. The quadratic model was developed and tested using the same training and testing sets which were employed to define the linear model above-discussed. In this case the procedure of feature selection reduced the variables from 78 to only 3 features, thus the number of parameters to be estimated from the training set was equal to 18. For the quadratic model the AUC values were equal to 0.780 and 0.785 for training and testing sets, respectively, while the 95% confidence interval of AUC, calculated in the testing set, was equal to [0.738-0.832].

The above results indicated that both linear and quadratic Bayesian classifiers had acceptable discrimination capacities on test data, because their AUC was always greater that 0.7 and less than 0.8 [14]. However, the quadratic model showed a greater power of generalization, because the AUC values calculated in the training and testing sets were nearly identical. This may be due to two causes:

1. the quadratic model, which releases the assumption of identical covariance matrices, is better suited to the actual probabilistic structure of the problem,
2. the number of parameters to be estimated from the training set for the quadratic model is smaller than that of the linear model (18 vs. 52), thanks to the smaller number of features needed to define the model.

6. The naïve Bayes approach

Even if the Gaussian assumption is correct, when the number of features necessary to define the decision rule is high, the previous linear and quadratic Bayesian models can lead to unsatisfactory results, due to the large number of parameters that must be estimated from the

http://dx.doi.org/10.5772/52402

training data to identify the class-conditional probability functions. In these cases one way to try to solve the problem is to use a naïve Bayes approach.

The naïve Bayes approach assumes that the features are all conditionally independent of one another within each class [15]. Consequently, given a q-dimensional observation vector **x** = (x_1, x_2,….., x_q) and two classes ω_1 and ω_2, the *a posteriori* probabilities of equation 2 can be rewritten as

$$P(\omega_i \mid \mathbf{x}) = \frac{P(\omega_i)\prod_{j=1}^{q} p(x_j \mid \omega_i)}{p(\mathbf{x})} \qquad i = 1,2 \tag{10}$$

where $p(x_j \mid \omega_i)$ is the class-conditional probability of feature x_j (j=1,2,….,q) conditioned on ω_i.

This assumption dramatically simplifies the problem of estimation from training data. In fact one-dimensional class-conditional distributions can be estimated for each feature individually, reducing a multidimensional task to a number of one-dimensional tasks. Despite its simplicity, the naïve Bayes classifier can often outperform more sophisticated classification methods and shows good average performance in terms of classification accuracy, even when the assumption of independence does not hold [16-18]. Actually, the naïve Bayes classifier estimates the parameters required for accurate classification using less training data than many other classifiers. This makes it particularly effective for datasets containing many features.

A naïve Bayes classifier is generally easy to understand and its induction is extremely fast if all features are discrete. In particular, for each class ω_i, a discrete feature x_j with K values (d_{1j}, d_{2j},….., d_{Kj}) can be simply characterized by the K-dimensional vector [$P(d_{1j} \mid \omega_i)$, $P(d_{2j} \mid \omega_i)$,….., $P(d_{Kj} \mid \omega_i)$] describing the class-conditional probability of this variable. In these conditions the class-conditional probability functions in equation 10 can be estimated by frequency counts from the training set.

A similar scheme may also be used when the feature set contains discrete and/or continuous variables. In this case, each continuous feature x_j can be discretized by grouping the observations of the whole training set into M groups of equal size (quantile intervals), denoted Q_{1j}, Q_{2j},….., Q_{Mj}. Thus, the continuous feature x_j is also characterized by class-conditional probability vectors [$P(Q_{1j} \mid \omega_i)$, $P(Q_{2j} \mid \omega_i)$,….., $P(Q_{Mj} \mid \omega_i)$] where $P(Q_{hj} \mid \omega_i)$ (h = 1,2,….,M) is the probability that x_j lies in the quantile interval Q_{hj} conditioned on ω_i [i = 1,2].

Recently, a naïve Bayes classifier was used to develop a locally-customized model for planning transfusion requirements in cardiac surgery [19]. The procedure allows transfusion planning to be designed and customized to a specific clinical setting on the basis of local available evidence. The size of the sample used in the study was 3182 patients.

The available patient sample was divided into two classes:

a. class of *negative* patients: patients who received no red blood cells in the intensive care unit after cardiac surgery (1107 patients),

b. class of *positive* patients: patients who received at least one pack of red blood cells (2075 patients).

Firstly, eleven dichotomous variables and six continuous variables were chosen as a set of likely independent predictors for planning transfusion quantities on the basis of qualitative evidence and previous knowledge. Univariate statistical analysis was then performed for each potential predictor in order to assess statistical differences between the classes of patients. Two dichotomous variables did not show statistically significant differences between classes and were eliminated from classifier design. Therefore fifteen variables giving statistical differences between the classes were used as features to design the naïve Bayes classifier.

Before developing the naïve Bayes classifier, the six continuous features were rendered discrete by dividing each into ten intervals of equal frequency. This was done by identifying the deciles for the whole sample of patients.

In order to evaluate the performance of this decision model, a confusion matrix was generated by the leave-one-out method, according to which each case in the training set is analysed by a classifier trained using the rest of the cases (Table 1). The overall classification accuracy was 73.7% and its corresponding 99% confidence interval was [71.7%, 75.7%].

The results shown in Table 1 reveal that although the naïve Bayes classifier cannot exactly distinguish negative from positive patients, it correctly classified about three-quarters of cases with a SE of 71.2% and a SP of 78.4%. Thus it seems to provide useful additional information for recognizing patients with transfusion requirements.

		Predicted class		**Correct classification percentage**
		Positive	Negative	
Actual class	Positive	1478	597	71.2% (SE)
	Negative	239	868	78.4% (SP)
Overall correct-classification percentage: 73.7% 99% confidence interval: [71.7%, 75.7%]				

Table 1. Confusion matrix and correct-classification percentages obtained by the leave-one-out method. SE = Sensitivity, SP = Specificity

7. Naïve Bayes classifiers and scoring systems

Part of the difficulty of medicine is to turn qualitative judgments into quantitative assessments. To face this problem, physicians have often proposed scoring systems with the intent

http://dx.doi.org/10.5772/52402

to summarize a set of items by means of a quantitative score. While scoring systems should never be used in place of a physician's judgment, they are undoubtedly very simple and attractive tools to analyse data and to use as decision-support systems.

The idea of using scoring systems in public health issues goes back many years ago. For example, in 1911, Hill suggested an elementary score system for determining the real relative importance of the different infectious diseases [20].

At first the methodology used to develop scoring systems was rather empirical, but over time the design of such systems has made use of more reliable quantitative techniques. Although different methodologies can be used, a naïve Bayes approach can represent a straightforward way to develop trustworthy scoring systems. In fact, in the presence of discrete (qualitative or quantitative) variables and discretized continuous variables, equation 10 can be rewritten as

$$P(\omega_i \mid \mathbf{x}) = \frac{P(\omega_i)\prod_{j=1}^{q1} P(d_{xj} \mid \omega_i)\prod_{j=1}^{q2} P(Q_{xj} \mid \omega_i)}{P(\mathbf{x})} \qquad i = 1,2 \tag{11}$$

where q1 and q2 are the number of discrete and continuous features, respectively, (q1 + q2 = q), d_{xj} is the value of the j-th discrete variable and Q_{xj} the quantile interval containing the j-th continuous variable.

The decision rule of equation 1 becomes

$$\begin{aligned} &\frac{P(\omega_1)}{P(\omega_2)}\prod_{j=1}^{q1}\frac{P(d_{xj} \mid \omega_1)}{P(d_{xj} \mid \omega_2)}\prod_{j=1}^{q2}\frac{P(Q_{xj} \mid \omega_1)}{P(Q_{xj} \mid \omega_2)} \geq 1 \quad \Rightarrow \quad \mathbf{x} \in \omega_1 \\ &\frac{P(\omega_1)}{P(\omega_2)}\prod_{j=1}^{q1}\frac{P(d_{xj} \mid \omega_1)}{P(d_{xj} \mid \omega_2)}\prod_{j=1}^{q2}\frac{P(Q_{xj} \mid \omega_1)}{P(Q_{xj} \mid \omega_2)} < 1 \quad \Rightarrow \quad \mathbf{x} \in \omega_2 \end{aligned} \tag{12}$$

Writing equation 12 in logarithmic form, the decision rule is

$$\begin{aligned} &S = \sum_{j=1}^{q1}\ln\frac{P(d_{xj} \mid \omega_1)}{P(d_{xj} \mid \omega_2)} + \sum_{j=1}^{q2}\ln\frac{P(Q_{xj} \mid \omega_1)}{P(Q_{xj} \mid \omega_2)} \geq \ln\frac{P(\omega_2)}{P(\omega_1)} \quad \Rightarrow \quad \mathbf{x} \in \omega_1 \\ &S = \sum_{j=1}^{q1}\ln\frac{P(d_{xj} \mid \omega_1)}{P(d_{xj} \mid \omega_2)} + \sum_{j=1}^{q2}\ln\frac{P(Q_{xj} \mid \omega_1)}{P(Q_{xj} \mid \omega_2)} < \ln\frac{P(\omega_2)}{P(\omega_1)} \quad \Rightarrow \quad \mathbf{x} \in \omega_2 \end{aligned} \tag{13}$$

We can note that S is a sum of q addends

$$S = \sum_{j=1}^{q} w_{xj} \tag{14}$$

where w_{xj} (j = 1,2,….,q) can be regarded as terms separately weighing the information given by each feature x_j. These weights are log-likelihood ratios, which can be determined from the training set. S is the total score: if S is less than ln [$P(\omega_2)/P(\omega_1)$], the observation vector **x** is classified in ω_2, otherwise **x** is classified in ω_1. ln [$P(\omega_2)/P(\omega_1)$] is the threshold value, independent of **x**, which is equal to 0 when $P(\omega_1) = P(\omega_2)$.

In conclusion, the decision rule in equation 13 defines a scoring system corresponding to a naïve Bayes classifier.

8. Bayesian networks

The simplicity and surprisingly high accuracy of the naïve Bayes classifier have led to its wide use, and to many attempts to extend it. In particular, naïve Bayes is a special case of a Bayesian network.

Bayesian networks are a flexible tool particularly suited to problem-solving [21]. They are graphic probability models of knowledge that lend themselves to modelling in situations of uncertainty. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Thus, given symptoms, the Bayesian network can be used to compute the probabilities of the presence of different diseases.

A Bayesian network is a directed acyclic graph, in which nodes represent random variables and arcs represent conditional dependencies between variables. Each nodes is associated with a probability table. The probability of a node can be calculated when the values of its incoming nodes are known. To describe a Bayesian network we need to specify the graph structure and the values of each probability table based on data.

The example in figure 3 shows a simple Bayesian network in which the various nodes are represented as circles and the corresponding variables are dichotomous and therefore discrete. The arc between nodes A and C with arrow in direction A to C, indicates that A has an influence on C (A is parent of C). The lack of an arc between two nodes indicates their conditional independence: for example, nodes C and D are independent of each other.

Each node has quantitative probability information. Nodes with parents (nodes C and D in figure 3) are characterised by a conditional probability table that defines the effects of the parents on the node, whereas nodes without parents (nodes A and B) are associated with a probability *a priori*. Thus P(A) is the probability of event A and $x_{\bar{A}B}$ is the probability of event C if event B occurs but not event A. Once the topology of the Bayesian network has been determined, the conditional probability distribution of all variables and their parents must be specified.

http://dx.doi.org/10.5772/52402

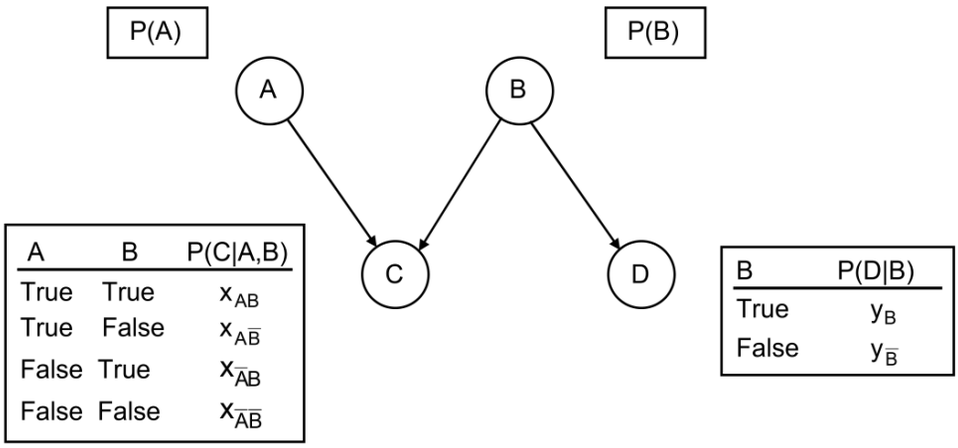

Figure 3. Simple example of Bayesian network with four nodes, two of which (nodes C and D) have parents.

The procedures required to develop and use models based on Bayesian networks may be not simple, but there is a number of commercial software packages. Information about various software packages available for Bayesian networks is provided by Murphy [22].

Learning the structure of a Bayesian network from data can be a critical point. An alternative approach can be to combine statements about the structure of Bayesian network from multiple experts into a single structure that more closely captures the dependencies in the domain. This structure is then refined, and its parameters estimated, using standard Bayesian network learning algorithms.

Bayesian networks can be applied to study major public health problems [23-25]. For illustrative purposes, we will start with a brief description of a fairly complex Bayesian network that has been proposed as a model of heart disease by researchers at the University of South Carolina [26]. Although, as the authors point out, the model is merely a prototype, it lends itself well to being used to explain the operation of a Bayesian network. The Bayesian network structure is shown in figure 4.

The structure proposed shows that the heart disease has five parents (adverse medicine, high blood pressure, atherosclerosis, family history and serum selenium). Two of these (adverse medicine and family history) have no parents, while the remaining three (high blood pressure, atherosclerosis and serum selenium) are, in turn, nodes with parents. In particular, atherosclerosis has four parents, three of which (high serum triglycerides, high serum LDL cholesterol and cholesterol/HDL) are influenced by the type of diet. Absence of moderate exercise directly influences the atherosclerosis. The structure in figure 4 also shows that heart disease influences four variables (ECG, angina pectoris, myocardial infarction and heart beat rate).

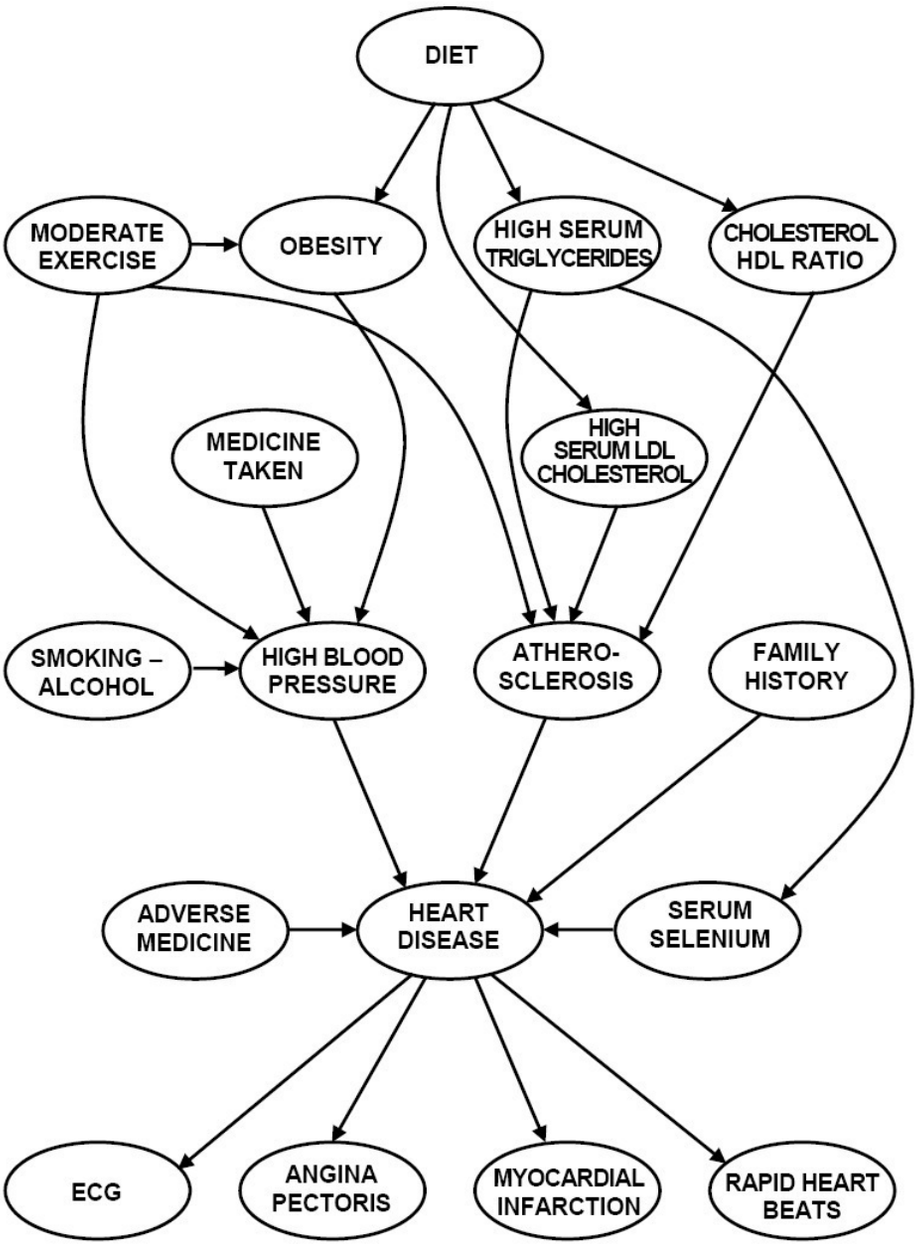

Figure 4. Bayesian network structure of the model of heart disease.

The above analysis indicates that the Bayesian network structure alone can be seen as a kind of mental map, i.e. as a diagram in which concepts are presented in graphic form. Of course, the diagram provides only qualitative information, but, by associating each node with a probability table, it is possible to obtain precise quantitative information on the whole system operation. For example, on the basis of their data, the authors found that the conditional probability of atherosclerosis is equal to 0.84 when moderate exercise is absent and serum triglycerides, serum LDL cholesterol and cholesterol HDL ratio are high [26]. The condition-

http://dx.doi.org/10.5772/52402

al probability of atherosclerosis drops to 0.20 when moderate exercise is present and serum triglycerides, serum LDL cholesterol and cholesterol HDL ratio are low [26].

Actually heart disease is a very complex system and the network shown above, despite its eighteen nodes, is a very simplified model that does not take into account important factors. For example, there is a significant chance that some inhibitor prevents atherosclerosis when serum triglycerides are high. Therefore the network in figure 4 is a useful example for explaining the operation of a Bayesian network, but does not lend itself to assess the actual strength of this approach.

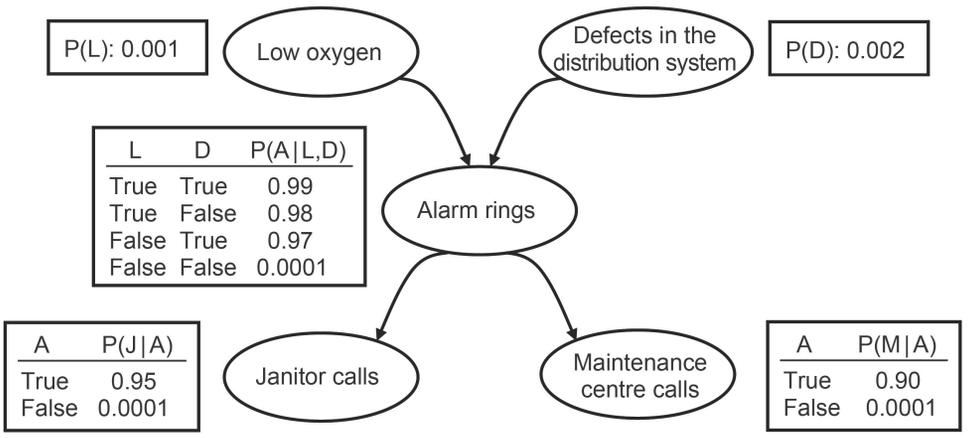

Figure 5. Bayesian network representing the medical gas plant control system. Events L, D, A, J and M represent low oxygen, defects in the distribution system, alarm rings, janitor calls and maintenance centre calls.

The strength of the approach is well illustrated by the Bayesian network which has been developed at the University Hospital of Siena (Italy) for assessing the performance of a control system of the medical gas plant (figure 5). The network takes into consideration a medical gas system connected to an alarm. The alarm rings when the pressure of oxygen falls below safety level in the main tank, but it may also ring for defects in the distribution system, such as a leaking pipe.

The alarm rings in two places, the janitor's lodge and the maintenance centre, which are responsible for calling the crisis unit. Both these places may commit errors in activating the crisis unit. For example, the janitor may not hear the alarm because he is busy giving information to users, or the maintenance centre may not hear it because of boiler noise. The network is a representation of a real situation where uncertainty exists. In other words, the network does not directly model the fact that the janitor is busy giving information to users or that the maintenance centre is noisy. These factors (and an infinite number of others) are summarised by the uncertainty associated with the links between the nodes "alarm rings" and "janitor calls" and between "alarm rings" and "maintenance centre calls". The probabil-

ity tables therefore enable an enormous set of circumstances to be summarised by an approximate model of a much more complex situation.

Once the Bayesian network and the probability tables have been specified, it is possible to evaluate the effective reliability of the hospital system in monitoring the medical gas plant, allowing answers to be given to questions about its actual functioning. For example, important questions to be answered can be:

a. What is the probability that the crisis unit is called by at least one of the two places where the alarm rings when the oxygen in the tank is insufficient or when the distribution system is damaged?

b. What is the probability that the oxygen system has neither of these problems when the crisis unit is called by the janitor (false alarm by janitor)?

c. What is the probability that the oxygen is insufficient or the oxygen system damaged when the janitor calls the crisis unit?

When the calculations are done, it emerges that the probability that the crisis unit be called when oxygen is low is 97.5%. Similarly, the probability that the janitor or the maintenance centre call the crisis unit for damage to the system is 96.5%. Actually in this example the alarm system is not infallible, because when the problem is low oxygen pressure in the tank there is a 98% probability of the alarm ringing and that probability falls to 97% when the problem is a defect in the distribution system (see table in figure 5). Consequently the above results suggest that the protocol in place to call the crisis unit (janitor plus maintenance centre) is efficient.

However, when the performance of the protocol is further evaluated, some criticalities emerge. The model makes it possible to determine that once the janitor has called the crisis unit, there is a 6.5% probability of sufficient oxygen and no damage to the oxygen distribution system, which is no longer negligible. This important discovery is not immediately obvious from the design of the system and the reliability of the various components (alarm, janitor and maintenance centre) taken separately. The result is actually due to the fact that the absolute probabilities of low oxygen and system damage are very small (0.001 and 0.002, respectively), because these events occur very rarely. Assuming a very rare event, even if the probability that the janitor calls in the absence of alarm ring is almost zero (0.0001), the network points out that a call from the janitor does not mean certainty or almost certainty of a problem in the system. In other words, there is a non negligible probability (about 6.5%) of there being neither low oxygen nor damages when the janitor calls. This weakness is inherent to the system but does not have particularly serious consequences, because the only effect is that in some cases the crisis unit alerted by the janitor will not find any problem in the plant. Since this would rarely happen, its cost is very low.

The answer to the third question shows that once the janitor has called the crisis unit, the probability that oxygen is low in the main tank is about 31.5% and that the distribution system is damaged about 62%. This difference is due to the fact that while both events are rare, system damage has twice the probability a priori of low oxygen (0.002 vs. 0.001).

http://dx.doi.org/10.5772/52402

In summary, the response of the Bayesian network to the above questions indicates that in the case of problems, the protocol provides a very efficient response, practically equivalent to that of a pure alarm device, but a call from the janitor may sometimes be a false alarm.

9. Conclusion

The Bayesian approach is a way of formulating and dealing with problems where decisions have to be made under a state of uncertainty. This situation is very frequent in problems of medical decision-making.

Although Bayes's theorem was published in 1763, the use of Bayesian techniques to develop decision-support systems in medicine and health management is recent. The design of a reliable Bayesian model often requires the use of complex algorithms and its development may involve not trivial calculations. This presented a barrier to the use of the Bayesian approach until the development of numerical methods and powerful computers during the late 20th century.

The advantages of the Bayesian approach are that it offers intuitive and meaningful inferences, that it gives the ability to tackle complex problems and that it allows the incorporation of prior information in addition to the data. In particular, the simple and intuitive nature of the Bayes's theorem as a mechanism for synthesising information and updating personal beliefs is an attractive feature, which has facilitated its spread in the medical field.

In conclusion, although the Bayesian approach is not the only way to tackle the problem of decision making under uncertainty, it is undoubtedly a very useful tool to be used in medicine and health management, because it allows the professional to make decisions on rational bases, thus making the choices easier to explain and defend.

Acknowledgements

This work was partly financed by the University of Siena, Italy.

Author details

Emanuela Barbini[1], Pietro Manzi[1] and Paolo Barbini[2,3*]

*Address all correspondence to: paolo.barbini@unisi.it

1 Health-Care Management, University Hospital of Siena, Italy

2 Department of Surgery and Bioengineering, University of Siena, Italy

3 Biomedical Engineering Unit, University Hospital of Siena, Italy

References

[1] Kadane, J. B. (2005). Bayesian methods for health-related decision making. *Statistics in Medicine*, 24(4), 563-567.

[2] Sheppard, J. W., & Kaufman, M. A. (2005). A Bayesian approach to diagnosis and prognosis using built-in test. *IEEE Transactions on Instrumentation and Measurement*, 54(3), 1003-1018.

[3] Barbini, E., Cevenini, G., Scolletta, S., Biagioli, B., Giomarelli, P., & Barbini, P. (2007). A comparative analysis of predictive models of morbidity in intensive care unit after cardiac surgery- Part I: model planning. BMC Medical Informatics and Decision Making 7:35 10.1186/1472-6947-7-35 published online 22 November 2007

[4] Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (2000). Bayesian methods in health technology assessment: a review. *Health Technology Assessment*, 4(38), 1-130.

[5] Berry, D. A., & Stangl, D. K. (1996). Bayesian biostatistics. *New York: Marcel Dekker, Inc*.

[6] Broemeling, L. D. (2007). Bayesian biostatistics and diagnostic medicine. *Boca Raton, FL: Chapman & Hall/CRC*.

[7] Moyé, L. A. (2008). Elementary Bayesian biostatistics. *Boca Raton, FL: Chapman & Hall/CRC*.

[8] Bayes, T. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. (1763), *Philosophical Transactions, Giving Some Account of the Present Undertakings, Studies and Labours of the Ingenious in Many Considerable Parts of the World*, 53, 370-418.

[9] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. *New York: John Wiley & Sons, Inc*.

[10] Friedman, G. D. (2004). Primer of epidemiology. *New York: The McGraw-Hill Companies, Inc*.

[11] Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, 38(5), 404-415.

[12] Fukunaga, K. (1990). Introduction to statistical pattern recognition. *San Diego, CA: Academic Press*.

[13] Cevenini, G., Barbini, E., Scolletta, S., Biagioli, B., Giomarelli, P., & Barbini, P. (2007). A comparative analysis of predictive models of morbidity in intensive care unit after cardiac surgery- Part II: an illustrative example. BMC Medical Informatics and Decision Making 7:36 10.1186/1472-6947-7-36 published online 22 November 2007

[14] Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression. *New York: John Wiley & Sons, Inc*.

[15] Mitchell, T. M. (1997). Machine Learning. *New York: The McGraw-Hill Companies, Inc*.

[16] Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3), 103-130.

[17] Lavrac, N., Zupan, B., Kononenko, I., Kukar, M., & Keravnou, E. T. (1998). Intelligent data analysis for medical diagnosis: using machine learning and temporal abstraction. *AI Communications*, 11(3-4), 191-218.

[18] Demichelis, F., Magni, P., Piergiorgi, P., Rubin, M. A., & Bellazzi, R. (2006). A hierarchical naïve Bayes model for handling sample heterogeneity in classification problems: an application to tissue microarrays. BMC Bioinformatics 7:514 10.1186/1471-2105-7-514 published online 24 November 2006

[19] Cevenini, G., Barbini, E., Massai, M. R., & Barbini, P. (2011). A naïve Bayes classifier for planning transfusion requirements in heart surgery. *Journal of Evaluation in Clinical Practice.*, Article first published online: 23 Aug, 10.1111/j.1365-2753.2011.01762.x.

[20] Hill, H. W. (1911). A "score system" for determining the real relative importance of the different infectious diseases. *Journal of the American Public Health Association*, 1(1), 7-9.

[21] Russell, S., & Norvig, P. (2010). Artificial intelligence: a modern approach. *Upper Saddle River, NJ: Pearson Education*.

[22] Murphy, K. (2012). Software packages for graphical models/Bayesian networks. http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html, update 14 June.

[23] Burnside, E. S., Rubin, D. L., Fine, J. P., Shachter, R. D., Sisney, G. A., & Leung, W. K. (2006). Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. *Radiology*, 240(3), 666-673.

[24] Li, J., Shi, J., & Satz, D. (2008). Modeling and analysis of disease and risk factors through learning Bayesian networks from observational data. *Quality and Reliability Engineering International*, 24(3), 291-302.

[25] Mnatsakanyan, Z. R., Burkom, H. S., Coberly, J. S., & Lombardo, J. S. (2009). Bayesian information fusion networks for biosurveillance applications. *Journal of the American Medical Informatics Association*, 16(6), 855-863.

[26] Ghosh, J. K., & Valtorta, M. (2000). Building a Bayesian network model of heart disease. *Proceedings of the 38th annual on Southeast regional conference, ACM-SE00 ACM-SE 2000- 38th Annual ACM Southeast Regional Conference, 7-8 April 2000, Clemson, South Carolina, USA. New York, NY, USA: ACM*, 2000, 239-240, 10.1145/1127716.1127770.

IntechOpen