On Leukocytes Classification: a Comparative Study

Verónica Rodríguez-López and Raúl Cruz-Barbosa

Computer Science Institute
Universidad Tecnológica de la Mixteca
69000, Huajuapan, Oaxaca, México
{veromix,rcruz}@mixteco.utm.mx

(Paper received on November 28, 2010, accepted on January 28, 2011)

Abstract. Leukocytes classification is a complex task, where the main problems are due to morphological diversity between cells of the same type and similar features found in different types of cells. In this paper a comparative study, in terms of classification accuracy, of Bayesian networks and neural networks for leukocyte classification is presented. The design of two Bayesian network models based on the expert's knowledge and data, a naive Bayes model and a multilayer perceptron neural network are presented. The experimental results have shown that a simple naive Bayes model is a suitable classifier for this task.

Keywords: Bayesian networks; neural networks; classification; leukocytes.

1 Introduction

White blood cells, or leukocytes, are cells of the immune system involved in defending the body against infection. There are five types of leukocytes that normally appear in blood: neutrophils, basophils, eosinophils, lymphocytes, and monocytes [7].

One of the most common requested test in a hematology laboratory is a complete blood count (CBC). As part of the CBC, a white blood cell count and a differential white blood cell count are done. The former measures the total number of white blood cells in a volume of blood given. The latter consists of a blood examination to determine the presence and the number of different types of white blood cells. The total number and the proportion of each type of leukocytes are associated with a person's health status [6, 3].

Leukocytes can be counted by either manual or automated hematology analyzers. The manual leukocytes count is a time consuming task, and highly dependent on lab technician skills who performs the differential analysis. Human classification errors are the main source of misclassification in the manual counts, where the main problem is the scarcity of cell samples (usually, sample sizes range from 100 to 200). On the other hand, automated hematology analyzers classify cell populations using both electrical and optical techniques. These machines decrease the time of performing routine examinations and at the same time increase cells classification accuracy. However, these analyzers are unable to accurately identify and classify all types of cells and are, particularly, insensitive to abnormal or immature cells. For this reason, most tests performed by these equipments will require a review of a skilled lab technician for definitive cell type identification [7].

To help lab technicians on leukocytes identification, many computational systems based on digital image processing and pattern recognition techniques have been developed. Despite several systems have reported a good performance [5, 9, 14], automation of leukocytes recognition is not an easy task. There are two main problems in this process. Firstly, morphological diversity is found between cells of the same type (e.g. neutrophil morphology). Secondly, similar features as shape and texture are found in different types of cell. These problems are illustrated in Figs. 1 and 2, respectively.

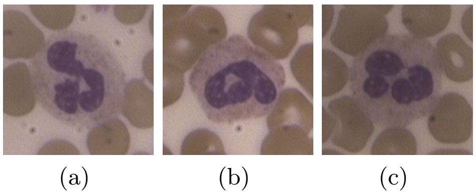

Fig. 1. Morphological diversity of neutrophil cell is presented in (a), (b) and (c).

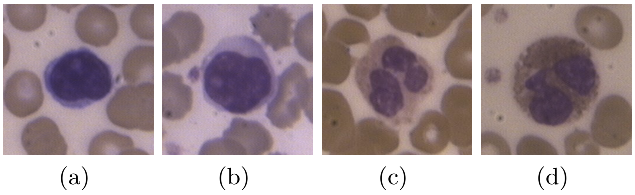

Fig. 2. Similar features between different types of leukocytes: (a) and (b) cells correspond to a lymphocyte and monocyte cell, respectively; (c) and (d) correspond to a neutrophil and eosinophil cell, respectively.

In this work, we compare the performance, in terms of classification accuracy, of Bayesian networks and neural networks for discrimination of five types of leukocytes. On the one hand, Bayesian networks have demonstrated to be useful as both a classifier and a powerful tool for knowledge representation and inference under conditions of uncertainty [8]. On the other hand, neural networks are a promising alternative to various conventional classification methods. These networks are data driven self-adaptive methods and they can adjust themselves to data without any explicit specification of functional or distributional form for the underlying model [16].

This paper is organized as follows. In Sect. 2, a brief description about Bayesian networks and neural networks is presented. The description of the Bayesian network models and neural network model design for leukocytes classification and the corresponding results are shown in Sect. 3. Finally, the conclusions are presented in Sect. 4.

2 Background

2.1 Bayesian networks

Bayesian networks (BN), also known as belief networks, belong to the probabilistic graphical models family. These graphical structures are used for knowledge representation of uncertain domains and when they work with statistical techniques together present several advantages for data analysis [8].

A formal definition of a BN is as follows. A Bayesian network model, or simply a Bayesian network, is a pair (D, P), where D is a directed acyclic graph (DAG), $P = \{p(x_1|\pi_1), ..., p(x_n|\pi_n)\}$ is a set of n conditional probability distributions, one for each variable, and Π_i is the set of parents of node X_i in D [4]. The set P defines the associated joint probability distribution as

$$p(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(x_i|\pi_i) \tag{1}$$

The simplest form of a Bayesian network is the naive Bayes model, in which the root node of a tree-like structure corresponds to a class variable. Also, this node is the only one parent for each attribute. The key assumption of the naive Bayes model is that all attributes are independent given the value of the class variable. Using this assumption, the conditional probability distribution for the class variable is very easy to calculate.

The naive Bayes assumption is helpful when we face high dimensionality input spaces. It is also useful when input vectors contain both discrete and continuous variables, since each one can be represented separately using appropriate models (e.g., Bernoulli distributions for binary observations or Gaussians for real-valued variables) [2].

Naive Bayes has been used as a simple and effective classifier in the Pattern Recognition field. It has two advantages over many other classifiers. Firstly, it is easy to construct, as the structure is given a priori. Secondly, a very efficient classification process is obtained.

2.2 Neural networks

Artificial neural networks, or simply neural networks, theory is an attempt at modeling the information processing capabilities of nervous systems. In mathematical terms, a neural network model is defined as a directed graph with the following properties:

1. A state variable n_i is associated with each node (neuron) i.
2. A real-valued weight w_{ik} is associated with each link (ik) between two nodes i and k.
3. A real-valued bias θ_i is associated with each node i.
4. A transfer function $f_i(n_k, w_{ik}, \theta_i, (k \neq i))$ is defined for each node i, which determines the state of the node as a function of its bias, weights (of its incoming links) and states of the nodes connected to it.

The transfer function usually takes the form $f(\sum_k w_{ik}n_k - \theta_i)$, where $f(\cdot)$ is a discontinuos step function or its smoothly increasing generalization known as sigmoidal function. Nodes without links toward them are called input neurons; output neurons are those with no link leading away from them [11].

The multilayer feedforward neural network, or equivalently referred to as multilayer perceptrons (MLP), is a very popular model in neural networks. A MLP has a layered structure: an input layer consisting of sensory nodes, one or more hidden layers of computational nodes, and an output layer that calculates the outputs of the network. The most common algorithm for training a MLP is named Backpropagation. In this algorithm the information is only propagated in the forward direction and there are no feedback loops. Even it does not have feed back connections, errors are back propagated during training. That is, the computations are passed forward from the input to the output layer, then the calculated errors are propagated back in the opposite direction to update the weights in order to obtain a better performance of the model [15].

3 Experiments

3.1 Experimental design and settings

The main objectives of the experiments are twofold. Firstly, we aim to explore the use of Bayesian network models for classifying all types (neutrophils, basophils, eosinophils, lymphocytes, and monocytes) of leukocytes. Secondly, a performance comparative study of Bayesian network models and neural network models is proposed.

Our experiments were carried out as follows. Initially, two tree structure Bayesian networks (TBN) models, which consider the expert's knowledge and medical literature, were designed. Then, we built a naive Bayes model and a MLP neural network model using the same features identified for the proposed Bayesian networks.

For the first experiment, the TBN-A and TBN-B models were designed. For simplicity, although perhaps there are other more suitables topologies, we used a tree structure in these models. For both models, we proposed a leukocyte classification node as the main one.

In the TBN-A model, we aimed to use some characteristics that experts take into account for the classification process. In accordance with the expert's knowledge these important characteristics are shape, size, and texture of nucleus as well as size and texture of cytoplasm. These characteristics are incorporated into the model as discrete latent variables (or discrete latent nodes) and are connected with the (classification) principal node. Furthermore, for the Bayesian network structure building we placed some observable nodes (which are linked to the latent variables) representing the description or measurements of the corresponding features (see Fig. 3). These measurements are obtained by application of digital image processing techniques. The observable nodes are continuous variables that have a normal distribution. The description of the incorporated knowledge into the TBN-A model is presented as follows.

The first characteristic considered into the TBN-A model is the shape of the nucleus. The nucleus shape of lymphocytes is round, and the monocytes shape have a great reniform or horseshoe-shaped nucleus. The nucleus of neutrophils have from 2 to 5 lobules, it can present S, C, or glass shapes. The nucleus of eosinophils have 2 lobules and usually it is glass shaped. The nucleus of basophils is bi- or tri-lobed, but it is hard to see because of the number of granules which hide it [3, 7, 6]. This knowledge about the shape of nucleus is encoded into the nucleus shape node. The estimation of this shape is obtained by means of region descriptors, particularly, the compactness, dispersion, and the first Hu moment [12] were used. These descriptors were included into the TBN-A model as compactness, dispersion, and MH1 nodes.

Since nucleus size is more relevant than cytoplasm size for leukocytes identification, only the nucleus size is considered for the TBN-A model. Then, the nucleus size is measured as the ratio of number of pixels that belong to the corresponding region to the total number of pixels of the cell (nucleus and cytoplasm pixels). This nucleus size information is included into the nucleus size node, which was linked with the nucleus shape node due to the relationship between these two features.

The cytoplasm texture is an important characteristic of leukocytes, since it allows to group the cells by the presence or absence of granules in their cytoplasm [7]. The granulocyte type cells are neutrophils, basophils, and eosinophils. The agranulocyte cells are lymphocytes and monocytes. In order to get information about the cytoplasm texture, the energy descriptor [12] is used. This knowledge about the cytoplasm texture and its corresponding descriptor are captured with the cytoplasm texture and energyC nodes.

The texture of nucleus is another important characteristic of leukocytes that is reported in medical literature [7, 6]. For this reason, we included this knowledge into the TBN-A model in a similar way as the cytoplasm texture is.

The colour of cytoplasm is the last feature of leukocytes taken into account for the TBN-A model. The granulocyte leukocytes are characterized by the presence of differently staining colour granules in their cytoplasm: neutrophils have pink colour granules, eosinophils have orange granules, and basophils have dark purple granules. For the agranulocyte cases, the cytoplasm colour for lymphocytes is light blue and for monocytes is greyish blue [3, 6]. The colour descriptor is obtained through the average intensity value using the RGB space. The knowledge about the colour is encoded into the cytoplasm colour, Rvalue, Gvalue, and

Bvalue nodes. In summary, the topology of the TBN-A model is shown in Fig. 3.

In the second part of the first experiment, we explored the possibility to find a tree type Bayesian network model with a minimum set of nodes, which performs leukocyte classification with an acceptable degree of accuracy. A definition of the new model was found by modifying the TBN-A model. The modification is as follows. Analyzing the TBN-A model, we observed that the cytoplasm colour node is a redundant node because it does not encode uncertain information. For this reason, the cytoplasm colour node is removed. Since either cytoplasm or nucleus texture is described by one measurement we decided to remove the cytoplasm and nucleus texture nodes in the TBN-A model. We hypothesize that the energy's nodes are enough to consider the texture information. Following the previous observations, we defined the second Bayesian network model, namely TBN-B. The topology of the TBN-B model is presented in Fig. 4.

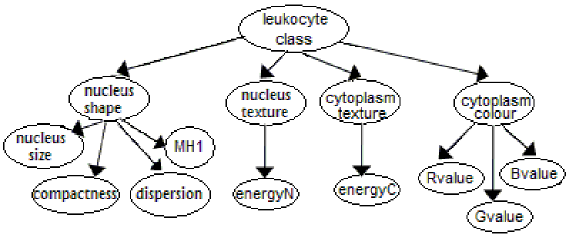

Fig. 3. Topology of the TBN-A model for leukocytes classification.

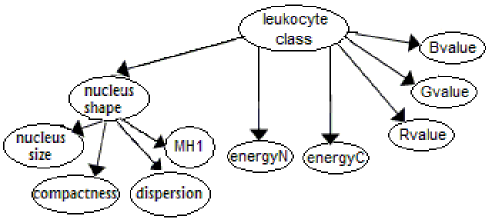

Fig. 4. Topology of the TBN-B model for leukocytes classification.

For the second experiment, the naive Bayes and the neural network models are built. To build both models, we used the most important features previously identified in the tree type Bayesian network models designed for leukocytes classification. These features are image descriptors (which are encoded in the observable or leaf nodes) used in the TBN-A and the TBN-B models.

Although a naive Bayes model does not consider domain knowledge and real dependence relationships between features, is a simpler model than other types of BN and has a good performance in classification problems. Analyzing the TBN-A and TBN-B models, we observed that there are no dependence relationships between the observable nodes, thus we hypothesize that a naive Bayes model could perform as well as a proposed BN model for leukocytes classification. Our naives Bayes model, namely NB, is showed in Fig. 5.

For the case of the MLP model, we built a neural network with a 9-N-5 topology, i.e. 9 input nodes, a hidden layer of N (where N will be determined in a range from 15 to 100 units) neurons, and a final output layer with 5 neurons providing the predicted leukocyte class. The input variables used in the first layer are the same input variables as in the NB model and the output layer corresponds with the five types of leukocytes (see Fig. 6). For training the constructed MLP, two faster algorithms than gradient descent are used: an heuristic-based and a numerical optimization-based method, namely, Resilient Backpropagation [13] and Scaled Conjugated Gradient [10], respectively.

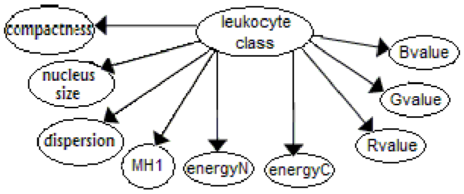

Fig. 5. Topology of the NB model for leukocytes classification.

3.2 Experimental results and discussion

In order to compare the performance, in terms of classification accuracy, of the Bayesian network and the neural network models, we used a set of 190 leukocytes colour images with a resolution of 256 × 256 pixels. The images were obtained by using a microscope that has an in-built CCD camera with a resolution of 640 × 480 pixels. The manual selection and cut of leukocytes region were applied to all images. For the nucleus and cytoplasm segmentation, we used a free software developed by Zoltan Kato [1]. The image set is formed by 8 basophils, 72 neutrophils, 9 eosinophils, 31 monocytes, and 70 lymphocytes. All images were previously classified by a human expert.

The classification performance of the designed Bayesian networks was evaluated by five-fold cross-validation. The parameters of the corresponding models

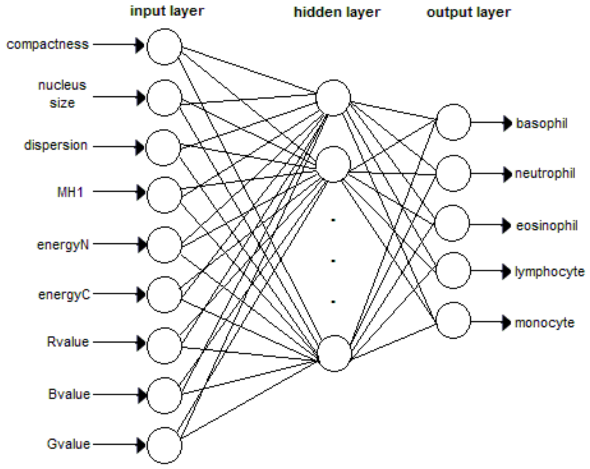

Fig. 6. Topology of the MLP model for leukocytes classification.

were obtained by using maximum likelihood estimation from complete data [8]. These models were tested using the Hugin Lite 7.3®software. The MLP model was trained and tested using the Matlab®software. As mentioned in the previous section, Resilient Backpropagation (RBP) and Scaled Conjugated Gradient (SCGBP) algorithms were used for training these neural network models. Here, the image set was divided into three sets: 60% for training set, 20% for validation, and 20% for independent test set.

The classification accuracy results for the proposed Bayesian network models and the MLP model, for the test set, are shown in Table 1. The results are better for the MLP (using 20 and 15 neurons in the hidden layer for RBP and SCGBP, respectively) and the NB models, in this case the models without domain knowledge. The domain knowledge was considered into the models with lower perfomance (the TBN-A and the TBN-B models). It seems that the domain knowledge implies dependencies between variables at same level, which were not considered into the proposed tree Bayesian network models. On the other hand, the results from Table 1 compare favourably with those classifiers that consider less types of leukocytes than the ones used here. For example, the classifiers presented in [9] and [5] consider only the most common types of leukocytes. In [9], they only classify four types (neutrophils, eosinophils, lymphocytes, and monocytes) of leukocytes achieving 86% of classification accuracy. Besides in [5], they classify neutrophils, eosinophils, and lymphocytes with 84% of accuracy.

The classification accuracy results for each type of leukocyte are presented in Table 2. Here, it can be observed that both Bayesian and neural network models can deal with the classification of all types of leukocytes, including basophils and eosinophils, which are, usually, imbalanced classes (they appear less often in blood cells). Also, the TBN-B network has achieved better results than the TBN-A model as we expected. Although the classification results for NB and MLP variants are the same (as shown in Table 1), it can be observed, from Table 2, these results are better balanced (for each class) in NB than in MLP variants.

In general, the experimental results have shown that the simplest Bayesian network, NB model, is more suitable for leukocytes classification. Even though the results for NB and MLP models are similar, the NB model is simpler, easily implementable and faster than a neural network classifier. Furthermore, the construction of neural networks is a complex task because there are no principled methods for network parameters selection. In contrast, a naive Bayes model has an easy and fast construction procedure without parameters tuning process.

On the other hand, despite the tree Bayesian network models do not seem to be ideal for leukocytes classification, their design helped us to select the feature set used in NB and MLP models. Moreover, it should be emphasized that these models are simple classifiers that can be easily understood and verified by experts.

Table 1. Classification accuracy results for TBN-A, TBN-B, NB, MLP-RBP, and MLP-SCGBP models.

Classifier model	**classif. acc.**
TBN-A	89.5%
TBN-B	90.5%
NB	94.7%
MLP-RBP	94.7%
MLP-SCGBP	94.7%

4 Conclusions

In this paper, two tree Bayesian network models, a naive Bayes model and a multilayer perceptron neural network model (trained with two different algorithms) were tested for leukocytes classification. Despite the analyzed data set has no enough images of some types of leukocytes (imbalanced classes), all proposed classifiers have achieved a good performance, which are comparable with those reported in literature. Our proposed models, particularly, naive Bayes, can classify all types of leukocytes, including the less frequent types, with a high degree of accuracy.

The experimental results have shown that the naive Bayes and the MLP models outperformed the tree Bayesian network models. Although this performance is similar for NB and MLP models, the results suggest that simple naive Bayes model should be preferred over the complex MLP model for leukocytes classification.

Table 2. Classification accuracy results for each type of leukocyte of the TBN-A, TBN-B, NB, MLP-RBP, and MLP-SCGBP models.

Classifier model	type of leukocyte	classif. acc.
TBN-A	basophils	93.3%
	neutrophils	95.3%
	eosinophils	83.3%
	monocytes	61.0%
	lymphocytes	88.0%
TBN-B	basophils	93.3%
	neutrophils	95.3%
	eosinophils	83.3%
	monocytes	82.7%
	lymphocytes	89.5%
NB	basophils	100%
	neutrophils	95.3%
	eosinophils	83.3%
	monocytes	95.5%
	lymphocytes	95.8%
MLP-RBP	basophils	50%
	neutrophils	100%
	eosinophils	100%
	monocytes	100%
	lymphocytes	100%
MLP-SCGBP	basophils	100%
	neutrophils	100%
	eosinophils	100%
	monocytes	66.7%
	lymphocytes	92.3%

As future work, the construction of other types of Bayesian network models such as tree- and Bayesian network-augmented naive-Bayes (TAN and BAN) for leukocyte classification is considered. These models could then be compared with our proposed NB model and other kind of classifier (as support vector machines).

References

1. Berthod, M., Kato, Z., Yu, S., Zerubia, J.: Bayesian image classification using markov random fields. Image and Vision Computing (14), 285–295 (1996)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
3. Carr, J.H., Rodak, B.F.: Clinical Hematology Atlas. Saunders, 2nd. edn. (2004)
4. Castillo, E., Gutierrez, J.M., Hadi, A.S.: Experts systems and Probabilistic Networks Models. Springer-Verlag (1997)
5. Colunga, M.C., Siordia, O.S., Maybank, S.J.: Leukocyte recognition using EM-algorithm. In: Aguirre, A.H., Borja, R.M., García, C.A.R. (eds.) MICAI '09: Proceedings of the 8th Mexican International Conference on Artificial Intelligence. pp. 545–555. Springer-Verlag (2009)

6. Estridge, B.H., Reynolds, A.P., Walters, N.J.: Basic Medical Laboratory Techniques. Delmar Cengage Learning, 4th. edn. (1999)
7. Greer, J.P., Foerster, J., Rodgers, G.M., Paraskevas, F., Glader, B., Arber, D.A., Means, R.T.: Wintrobe's Clinical Hematology, vol. 1. Lippincott Williams & Wilkins, 12th. edn. (2009)
8. Heckerman, D.: A tutorial on learning with bayesian networks. Tech. rep., Microsoft Research (1996)
9. Mircic, S., Jorgovanovic, N.: Automatic classification of leukocytes. Journal of Automatic Control 16(1), 29–32 (2006)
10. Moller, M.F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 6(4), 525–533 (1993)
11. Muller, B., Reinhardt, J., Strickland, M.T.: Neural networks: An Introduction. Springer, 2nd. edn. (1996)
12. Nixon, M.S., Aguado, A.S.: Feature Extraction & Image Processing. Academic Press, 2nd. edn. (2007)
13. Riedmiller, M., Braun, H.: A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: Proceedings of the IEEE International Conference on Neural Networks. pp. 586–591 (1993)
14. Rodrigues, P., Ferreira, M., Monteiro, J.: Segmentation and classification of leukocytes using neural networks: A generalization direction. In: Bhanu Prasad, S.M.P. (ed.) Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, pp. 373–396. Springer Berlin / Heidelberg (2008)
15. Tanga, H., Tan, K., Zhang, Y.: Neural networks: computational models and applications. Springer (2007)
16. Zhang, G.P.: Neural networks for classification: A survey. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and. Reviews 30(4), 451–462 (2000)