Feature Relevance in Bayesian Network Classifiers and Application to Image Event Recognition

Mohand Said Allili and Siham Bacha

Abstract

An important problem in Bayesian Network Classifiers (BNCs) is to discover relevant variables that can achieve optimal classification performance. We propose a method based on Bayesian inference for estimating and incorporating feature relevance in classification using BNCs. We empirically validate our method on an application to event recognition in natural images using object and scene information.

Introduction

Several classification problems such as text categorization and visual recognition make use of hundreds of features to describe instances of data. In the training phase, features are extracted from labeled instances of data and used to build classification models. In the prediction phase, labels of newly observed data are estimated through the trained model. However, the presence of redundant and/or noisy features can bias the classification at hand (Drugan et al. 2010; Guyon et al. 2003; Tang et al. 2014). For instance, the presence of repetitive words in a text document can bias topic categorization (Dasgupta et al. 2007). The same problem can rise in event recognition in images where a repetitive object can cause errors in event prediction (Wu et al. 2015).

To reduce the effect of noisy features in classification, several techniques have been proposed in the literature (Guyon et al. 2003). *Filter* techniques select features independently of the classification model by measuring criteria such as information gain and correlation analysis (Lefakis et al. 2014). *Wrapper* techniques perform a greedy search through the feature space to choose subsets of features using the classification model (Kohavi et al. 1997). Though wrappers are less biased than filters, they are computationally intensive. Despite their performance, filter and wrapper methods perform hard feature section, where a feature is discarded if deemed irrelevant, even though it may be useful when combined with other features (Guyon et al. 2003). To alleviate this issue, *embedded* methods have been proposed to use feature weighting instead of feature selection in classification. Embedded techniques encode feature relevance directly in model construction, and thus enjoy the advantages of filters and wrappers while making better use of data (Guyon et al. 2006).

Embedding feature relevance in model construction can usually lead to better generalisation performance (Ouyed et al. 2014, Tang et al. 2014). Embedding schemes consist generally in incorporating penalty terms in the objective function of the classifier to cause shrinkage of non-discriminative feature weights, thus decreasing their influence in classification. For example, penalty terms using the L_1 norm have been successfully applied in support vector machines to reduce the effect of noisy features (Guyon et al. 2006; Weston et al. 2003). Other approaches have used weighting schemes in discriminative classifiers such as neural networks and LASSO (least absolute shrinkage and selection) methods and obtained better performance for classification (Tang et al. 2014). Whereas several methods have been proposed for embedding feature relevance in discriminative classifiers, relatively fewer works exist for their generative counterparts. One can mention the recent works that have introduced feature weighting in the naive Bayes classifier to reduce the effect of redundant features (Allili et al. 2015, Drungan et al. 2010; Zaidi et al. 2013). These methods have demonstrated their good performance compared to using feature selection methods.

This paper is an extension of feature weighting for relevance embedding in Bayesian network classifiers (BNC). The approach is proposed in the context of an application to event recognition in image albums using object and scene features. Our proposed BNC structure incorporates feature relevance in the generative model where each feature is associated with a relevance variable encoding its influence in discriminating event classes. For instance, it is natural that a 'christmas tree' object will better help determining the event *Christmas* than the object 'chair'. In the same vein, 'a snowy mountain' scene can help discriminate a 'skiing' event more than a 'landscape' scene. As training data are provided in the form of labeled images, the relevance of each feature is estimated through Bayesian inference. To infer event categories of new albums, we combine object and scene features and use the maximum a posteriori probability (MAP) with feature relevance for prediction. Experiments on the challenging PEC dataset (Bossard et al. 2009) have demonstrated the performance of the proposed approach with comparison with state-of-the-art methods.

Feature relevance in Bayesian Network Classifiers

Suppose that we have a data classification problem with class labels represented by the variable Y and n data attributes represented by variables $X_1, ..., X_n$ (see Figure 1.(a) for illustration). Bayesian network classifiers (BNC) are generative model classifiers where the conditional probability of each attribute X_i given the class label Y, $p(X_i|Y)$, is learned from the training data. Classification is then done by applying the Bayes rule to compute the maximum a posteriori probability $p(Y|X)$ for a particular instance of data $X = (x_1, ..., x_n)$ (Friedman et al. 1997).

The computation of $p(Y|X)$ is achieved by making a strong independence assumption where the attributes $X_1, ..., X_n$ are assumed conditionally independent given the class label Y. Although this assumption seems somewhat unrealistic, the BNC has a good performance in general.

Note that all features are taken equally important for classification in BNCs. However, this does not reflect the way some applications should make use of data for prediction. In visual recognition, for example, detecting some parts of an object in an image is sometimes sufficient to recognize the category of the object without further parsing the rest of the image. Likewise, an event can be determined by a subset of objects (e.g., 'snow' and 'Christmas tree' can carry more information about the event 'Christmas' than 'sofa' or 'bottle' objects). On the other hand, the presence of too many instances of an object in an image (e.g. 'books') can bias the classification to an event occurring in a book store (e.g. 'book purchase'), even if the image is captured during a 'Birthday' event celebrated inside a living room of a house containing a wardrobe of 'books'.

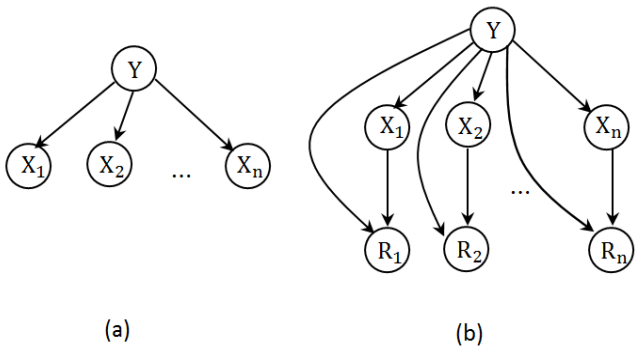

Figure 1: Graphical representation of a BNC: (a) without feature relevance, (b) with feature relevance, respectively.

To include feature relevance for classification, we propose an extension of the BNC structure as shown in Figure 1.(b). We suppose set of reevaluate parameters $R = \{R_1, ..., R_n\}$, where R_i encode the relevance of feature X_i for discriminating the different classes Y. Given the set of all variables $\mathcal{X}$ in the BNC, their joint probability is given by:

$$p(\mathcal{X}, Y, R) = p(Y) \prod_{i=1}^{n} p(X_i|Y)p(R_i|Y, X_i) \qquad (1)$$

The parameters of the different distribution constricting the factors of (1) can be estimated from labeled training data. In particular, the joint observation of class labels Y and values of the variables X_i can be a good indicator of the relevance of the variable X_i in discriminating the classes Y. This is shown in the context of the event recognition application developed in the next section.

Event recognition in images with BNC and feature relevance

Starting from the model proposed in Figure 1.(b), we propose a BNC structure for our application as shown in Figure 2, where classes consist of social event categories (e.g., 'hiking', 'cruise', 'wedding', 'birthday', etc.) in sets of images or albums. An album $\mathcal{A}$ can be constituted of N images $\{I_1, .., I_N\}$ and should be classified in one of the event categories. We suppose there are N_e events encoded by a discrete variable $e \in \{1, ..., N_e\}$ which follows a Multinoulli distribution $e \sim p(e|\eta)$, where η is its N_e-dimensional parameter vector. Images in an event category e can contain multiple instances of objects and scenes describing the semantic content and the context of the event. For example, a 'Christmas' event will contain 'humans', 'pieces of furniture', 'lights', etc., as objects and 'indoor house' as a scene.

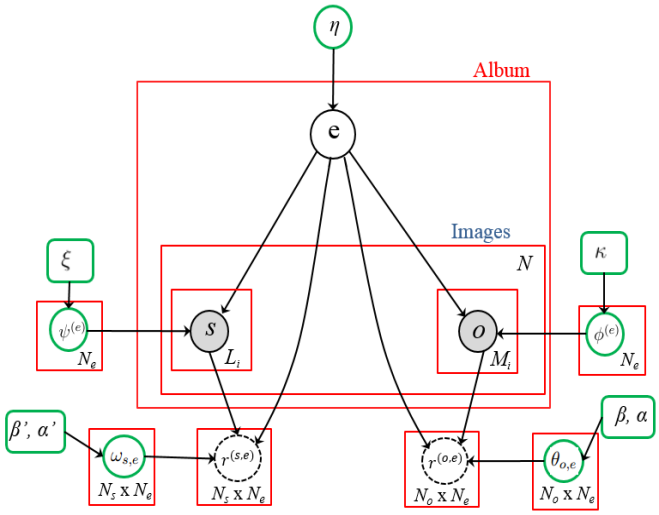

Figure 2: Detailed structure of the proposed BNC and feature relevance for event recognition in images.

The generative process of an image I_i in a given event category, then, will consist in generating scene and object instances. Let s and o be discrete random variables representing scene and object occurrences. We suppose an image I_i can contain up to L_i scene and M_i object instances denoted by $\mathbf{s}_i = \{s_{il}\}$, where $l \in \{1, \ldots, L_i\}$, and $\mathbf{o}_i = \{o_{im}\}$, where $m \in \{1, \ldots, M_i\}$.

Given an event category e, a scene instance s_{il} can be generated according to a Multinoulli distribution $Mut(\psi^{(e)})$ with a N_s-dimensional parameter vector $\psi^{(e)}$, where N_s is the number of scene categories. Likewise, an object instance

o_{im} can be generated according to another Multinoulli distribution $Mut(\psi^{(e)})$ with a N_o-dimensional parameter vector $\phi^{(e)}$, where N_o is the number of object categories. We assume the parameter vectors $\psi^{(e)}$ and $\phi^{(e)}$ have Dirichlet priors with hyper-parameters ξ and κ, respectively.

Finally, variables $r^{(o,e)}$ and $r^{(s,e)}$ encoding the relevance of each scene (resp. object) category with regard to event classes are generated through the learning data. First, the distribution of the relevance variables can be formulated as follows:

- Let $r^{(o,e)}$ be binary variable, where $r^{(o,e)} = 1$ if the object o is relevant to the event class e and $r^{(o,e)} = 0$, otherwise. We use a Bernoulli distribution $Ber(\theta_{o,e})$ to encode $p(r^{(o,e)} = 1|e, o) = \theta_{o,e}$, with $\theta_{o,e} \sim Beta(\alpha, \beta)$.
- Let $r^{(s,e)}$ be binary variable, where $r^{(s,e)} = 1$ if the scene s is relevant to the event class e and $r^{(s,e)} = 0$, otherwise. We use a Bernoulli distribution $Ber(\omega_{s,e})$ to encode $p(r^{(s,e)} = 1|e, s) = \omega_{s,e}$, with $\omega_{s,e} \sim Beta(\alpha', \beta')$.

The set of all parameters of the model is therefore $\Theta = \{\phi, \psi, \theta_{o,e}, \omega_{s,e}\}$. The learning of the elements of Θ is performed through Bayesian inference. Furthermore, we suppose that objects and scenes are independent given the event category. Thus, the object parameters $\{\phi^{(e)}, \theta_{o,e}\}$ and the scene parameters $\{\psi^{(e)}, \omega_{s,e}\}$ can be learned separately.

To estimate the relevance parameters $\theta_{o,e}$ and $\omega_{s,e}$, we use the maximum a posteriori of their probability. Without loss of generality, we presents the steps for estimating object relevance parameters. The same steps can be followed for estimating scene relevance parameters. The Bernoulli distribution for the variable $r^{(o,e)}$ is formulated as follows:

$$p(r^{(o,e)}|\theta_{o,e}) = (\theta_{o,e})^{r^{(o,e)}}(1 - \theta_{o,e})^{(1-r^{(o,e)})} \quad (2)$$

To estimate $\theta_{o,e}$, T samples are generated from the dataset. Let $\mathcal{D} = \{r_1^{(o,e)}, \ldots, r_T^{(o,e)}\}$ be T (calculated) observations for the variable $r^{(o,e)}$. Each observation is calculated by measuring the information gain of the presence/absence of objets in the occurence/non-occurrence of each event. For this purpose, the following lenient function is used:

$$r_i^{(o,e)} \approx 1 - \exp\left[-MI(x, e))\right], \quad (3)$$

where $MI(x, e)$ is the mutual information between object x and event e. Then, we estimate $\theta_{o,e}$ by maximizing its posterior probability:

$$\hat{\theta}_{o,e} = \arg\max_{\{\theta_{o,e}\}} \{p(\theta_{o,e})p(\mathcal{D}|\theta_{o,e})\} \quad (4)$$

where $p(\mathcal{D}|\theta_{x,e})$ is the likelihood of the parameters given by: $p(\mathcal{D}|\theta_{o,e}) = \theta_{o,e}^{N_1}(1 - \theta_{o,e})^{N_2}$, where $N_1 = \sum_i^T r_i^{(x,e)}$ and $N_2 = T - N_1$ give the number of samples where the object/scene is found relevant for the event class e according to Eq. (3). Having the $Beta(\alpha, \beta)$ prior for $\theta_{o,e}$ with hype-parameters (α, β) gives the following MAP estimation:

$$\begin{aligned} \hat{\theta}_{o,e} &= \arg\max_{\{\theta_{o,e}\}} (Beta(\alpha + N_1 - 1, \beta + N_2 - 1)) \\ &= \frac{\alpha + N_1 - 1}{\alpha + \beta + T - 2} \end{aligned} \quad (5)$$

Finally, the category of a new album $\mathcal{A}'$ containing N' images is inferred by maximizing the posterior probability of events given the album images: $\hat{e} = \arg\max_e p(e|\mathcal{A}', \Theta)$. Writing the album in term of images gives:

$$\begin{aligned} p(e|\mathcal{A}', \Theta) &= \prod_{i=1}^{N'} \left\{ \prod_{m=1}^{M_i} p(e|o_{im}, r^{(o,e)}, \Theta) \right. \\ &\quad \left. \prod_{l=1}^{L_i} p(e|s_{il}, r^{(s,e)}, \Theta) \right\}, \end{aligned} \quad (6)$$

where we assume the building blocks of an image I_i are constituted of scenes and objects: $I_i = \{\mathbf{o}_i, \mathbf{s}_i\}$, and the objects and scenes are independent given the class event.

Experimental results

We conducted experiments for validating the proposed model using the PEC (Personal Event Collections) dataset containing 61 000 images grouped into 140 albums. Each album is labeled by one of the following event categories: *Birthday, Children's birthday, Christmas, Concert, Cruise, Easter, Exhibition, Graduation, Halloween, Hiking, Roadtrip, Saint Patrick's day, Skiing and Wedding* (Bossard et al. 2013). The same experimental protocol suggested by (Bossard et al. 2013) is used for our evaluations, where 10 albums per class have been used for testing (140 albums in total). To learn the parameters of the model, we randomly selected six albums for each event class (84 albums in total) from the proposed training set. We used the Caffe toolbox (Jia et al. 2014) using GoogLeNet convolutional network architecture to detect contained objects/scenes in the images. We used the ImageNet dataset (Krizhevsky et al. 2014) and Places205 (Zhu et al. 2014) datasets to train the object and scene net detectors. For better efficiency, we resized the images to 256×256 pixels before feeding them to the object and scene networks.

We compared our approach incorporating relevance with another version not incorporating relevance in the BNC. These two versions are named R-OS-BNC and OS-BNC, respectively and their performance are shown in the first two columns of Table 1. Clearly, incorporating relevance has increased performance on average by 4.28%. We also compared our approach to recent methods proposed in the literature for event recognition in images (Bossard et al. 2013; Kwon et al. 2015; Tsai et al. 2011; Wu et al. 2015). Our method has yielded an average of 74.29%, exceeding the best average accuracies obtained by all the compared methods. More specifically, our method outperformed the others in events: *Birthday*, *Children's birthday*, *Easter*, *Graduation* and *Wedding*. For other event categories, we have achieved a close performance to the compared methods. In terms of the F_1 score, we have obtained an average score of 74.82%, exceeding the best F_1 scores obtained by (Wu et al. 2015), (Kwon et al. 2015), (Tsai et al. 2011) and (Bossard et al. 2013) by 17.17%, 36.2%, 14.71% and 18.66%, respectively.

Conclusions

We have proposed a BNC model structure incorporating feature relevance for classification. The relevance parame-

Events	OS-BNC	R-OS-BNC	Bossard et al. 2013	Wu et al. 2015	Tsai et al. 2011	Kwon et al. 2015
Birthday	20%	**30** %	10 %	12 %	0 %	0 %
Children's birthday	50%	**60** %	30 %	57 %	**60** %	10 %
Christmas	70%	80%	70 %	**89**%	60 %	40 %
Concert	100%	**100**%	**100**%	**100**%	80 %	100 %
Cruise	80%	80%	50%	**82**%	70 %	40 %
Easter	50%	**60**%	50 %	44%	**60** %	20%
Exhibition	70%	70 %	70%	**75**%	50 %	50%
Graduation	80%	**90**%	40%	69%	70 %	40%
Halloween	70%	70%	30%	**82**%	70 %	10%
Hiking	**80**%	**80**%	**80**%	52%	40 %	70%
Road trip	60%	60%	40%	**91**%	10 %	30%
Saint Patrick's day	60%	70%	30%	**98**%	40 %	90%
Skiing	**100**%	**100**%	100%	**100**%	**100** %	60%
Wedding	**90**%	**90**%	80%	77%	**90** %	10%
Average accuracy	70%	**74.28**%	55.71%	73.43%	57.14 %	41.71%
Average F_1 measure	71.01%	**74.82**%	56.16%	57.68%	60.11 %	38.62%

Table 1: Comparison of our method with state-of-art methods using the PEC dataset. Shown numbers are average precision values obtained for each event category.

ters are learned with to the other model parameters using Bayesian inference. The proposed approach has been successfully validated in the context of an application to event recognition in images. We have obtained better performance with regard to using BNC without incorporating feature relevance or using other methods for image event recognition based on discriminative classifiers.

References

Allili, M.S., Ziou D., 2015, Likelihood-Based Feature Relevance for Figure-Ground Segmentation in Images and Videos. *Neurocomputing*, 167: 658-670.

Bossard, L., Guillaumin, M., and Van Gool, L., 2013. Event Recognition in Photo Collections with a Stopwatch HMM. *IEEE ICCV*, 1193-1200.

Dasgupta, A., Drineas, P., Penn, B., Josifovski, V., and Mahoney, M.W., 2007. Feature Selection Methods for Text Classification. *ACM SIGKDD*, 230-239.

Drugan, M.M., and Wiering, M.A., 2010. Feature Selection for Bayesian Network Classifiers Using the MDL-FS Score. *Int'l J. of Approximate Reasoning*, 51(6):695-717.

Friedman, N., Geiger, D., and Goldszmidt, M., 1997. Bayesian Network Classifiers *ML*, 29:131163.

Guyon, I., and Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. *JMLR*, (3):1157-1182.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L., 2006. *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing, Springer.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T., 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *ACM ICM*, 675-678.

Kohavi, R., and John, G.H., 1997. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2):273-324.

Krizhevsky, A., Sutskever, I., and Hinton, G.E., 2012. ImageNet Classification With Deep Convolutional Neural Networks. *NIPS*, 1106-1114.

Kwon, H., Yun, K., Hoai, M., and Samaras, D., 2015. Recognizing Cultural Events in Images: A Study of Image Categorization Models. *IEEE CVPR Workshops*, 51-57.

Lefakis, L., and Fleuret, F., 2014. Jointly Informative Feature Selection. *AISTATS*, 567-575.

Ouyed, O., Allili, M.S., 2014, Feature Relevance for Kernel Logistic Regression and Application to Action Classification. *IEEE ICPR*, 1325-1329.

Tang, J., Alelyani, S., and Liu, H., 2014, *Data Classification: Algorithms and Applications*. Chapman and Hall.

Tsai, S.-F., Huang, T.S., and Tang, F., 2011. Album-Based Object-Centric Event Recognition. *IEEE ICME*, 1-6.

Weston, J., Elisseff, A., and Schölkopf, B., 2003 Use of the Zero-Norm with Linear Models and Kernel Models. *JMLR*, 3:1439-1461.

Wu, Z., Huang,Y., and Wang, L., 2015. Learning Representative Deep Features for Image Set Analysis. *IEEE Trans. on Multimedia*, 17(11):1960-1968.

Zaidi, N.A., Cerquides, J., Carman, M.J., and Webb, G.I, 2013. Alleviating Naive Bayes Attibute Weighting Independance Assumption by Attribute Weighting. *JMLR*, 14:1947-1988.