Data mining in Network engineering ——Bayesian Networks for Data Mining

Xiao Dan Wang [1]

*[1]Huanggu District Shenyang Liaoning China
[a]1308386584@qq.com

Keywords: Data mining; Bayesian networks; Network engineering; Bayesian approach; Probability distribution

Abstract. Nowadays,data mining is a hot topic in all sorts of fields. Potential science applications include, Telecommunications companies apply data mining to detect fraudulent network usage. Companies in many areas of business apply data mining to improve their marketing and advertising. Law enforcement uses data mining to detect various financial crimes. Given the well known complexity of Network engineering processes and artifacts, it is perhaps not surprising that data mining can be applied there as well. A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest [1]. When used in conjunction with statistical techniques, the graphical model has several advantages for data modeling. In this paper, discussing methods for constructing Bayesian networks from prior knowledge and summarize Bayesian statistical methods for using data to improve these models. With regard to the latter task, describing methods for learning both the parameters and structure of a Bayesian network.

Introduction

A Bayesian network has become a popular representation for encoding uncertain expert knowledge in expert systems. The techniques that have been developed are new and still evolving, but they have been shown to be remarkably effective for some data-modeling problems [2].

In this paper, we provide a tutorial on Bayesian networks and associated Bayesian techniques for data mining—the process of extracting knowledge from data. There are numerous representations available for data mining, including rule bases, decision trees, and artificial neural networks; and there are many techniques for data mining such as density estimation, classification, regression, and clustering [3,4]. So what do Bayesian networks and Bayesian methods have to offer? There are at least four answers [5]. One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for avoiding the overfitting of data.

The Bayesian approach to probability and statistics

To illustrate the Bayesian approach,consider a common thumbtack—one with a round, flat head that can be found in most supermarkets. If we throw the thumbtack up in the air, it will come to rest either on its point (heads) or on its head (tails). Suppose we flip the thumbtack N +1 times, making sure that the physical properties of the thumbtack and the conditions under which it is flipped remain stable over time. From the first N observations, we want to determine the probability of

heads on the N +1th toss.In the Bayesian approach, we assert that there is some physical probability of heads, but we encode our uncertainty about this physical probability using (Bayesian) probabilities, and use the rules of probability to compute our probability of heads on the N+1th toss.

We denote a variable by an upper—case letter (e.g., X, Y, Xi, θ), and the state or value of a corresponding variable by that same letter in lower case(e.g., X, Y, Xi, θ).To examine the Bayesian analysis of this problem, we denote a set of variables by a bold-face upper-case letter (e.g., X, Y, Xi). We use a corresponding bold—face lower-case letter (e.g., X, Y, Xi) to denote an assignment of state or value to each variable in a given set [6]. We say that variable set X is in configuration x. We use p(X=x|ξ) (or p(x|ξ) as a shorthand) to denote the probability that X=x of a person with state of information ξ. We also use p(x|ξ) the denote the probability distribution for X (both mass functions and density functions). Whether p(x|ξ) refers to a probability, a probability density, or a probability distribution will be clear from context. We use this notation for probability throughout the paper.

Returning to the thumbtack problem, we define θ to be a variable whose values θ correspond to the possible true values of the physical probability. We sometimes refer to θ as a parameter. We express the uncertainty about θ using the probability density function p(x|ξ). In addition, we use Xl to denote the variable representing the outcome of the l th flip, l = 1,…, N +1, and D = {X1=x1,…, XN = xN}to denote the set of our observations. Thus, in Bayesian terms, the thumbtack problem reduces to computing p(xN+1|D,ξ) from p(θ|ξ).

To do so, we first use Bayes' rule to obtain the probability distribution for θ given D and background knowledge ξ:

$$p(q \mid D,x) = \frac{p(q \mid x)q^{h}(1-q)^{t}}{p(D \mid x)} \tag{1}$$

where h and t are the number of heads and tails observed in D, respectively. The probability distributions p(θ|ξ) and p(θ|D,ξ) are commonly referred to as the prior and posterior for θ, respectively. The quantities h and t are said to be sufficient statistics for binomial sampling, because they provide a summarization of the data that is sufficient to compute the posterior from the prior. Finally, we average over the possible values of θ(using the expansion rule of probability) to determine the probability that the N+1th toss of the thumbtack will come up heads:

$$p(X_{N+1} = heads \mid D,x) = \int q\, p(q \mid D,x)dq = E_{P(q|D,x)}(q) \tag{2}$$

where EP(θ|D,ξ) (θ)denotes the expectation of θ with respect to the distribution p(θ|D,ξ).

To complete the Bayesian story for this example, we need a method to assess the prior distribution for θ. A common approach, usually adopted for convenience, is to assume that this distribution is a beta distribution:

$$p(q \mid x) = Beta(q \mid a_h, a_t)\frac{\Gamma(a)}{\Gamma(a_h)\Gamma(a_t)}q^{a_h-1}(1-q)^{a_t-1} \tag{3}$$

Where $\alpha_h>0$ and $\alpha_t>0$ are the parameters of the beta distribution, $\alpha=\alpha_h+\alpha_t$, and $\Gamma(\cdot)$ is the *Gamma* function which satisfies $\Gamma(x+1)= x\Gamma(x)$ *and* $\Gamma(1)=1$. The quantities α_h and α_t are often referred to as *hyperparameters* to distinguish them from the parameter θ. The hyperparameters α_h and α_t must be greater than zero so that the distribution can be normalized.

Regardless of the functional form, we can learn about the parameters given data using the Bayesian approach. As we have done in the binomial case, we define variables corresponding to the

unknown parameters, assign priors to these variables, and use Bayes' rule to update our beliefs about these parameters given data:

$$p(q \mid D,x) = \frac{p(D \mid q,x)p(q \mid x)}{p(D \mid x)} \tag{4}$$

In multinomial sampling, the observed variable X is discrete, having r possible states x1,…, xr. The likelihood function is given by

$$p(X = x_k \mid (\theta,\xi) = (\theta_k) \qquad k=1,\cdots,r$$

where $\theta = \{\theta 2,\ldots,\theta r\}$ are the parameters (The parameter $\theta 1$ is given by $1-\sum_{k=2}^{r}\theta_k$). In this case, as in the case of binomial sampling, the parameters correspond to physical probabilities. The sufficient statistics for data set D= {X1 = x1,…, XN = xN} is {N1,…, Nr}, where Ni is the number of times X = xk in D. The simple conjugate prior
used with multinomial sampling is the Dirichlet distribution:

$$p(\theta \mid \xi) = Dir(\theta \mid \alpha_1,\ldots,\alpha_r \equiv \frac{\Gamma(\alpha)}{\Pi_{k=1}^{r}\Gamma(\alpha_k)}\prod_{k=1}^{r}\theta_k^{a_k-1} \tag{5}$$

where $\alpha = \sum_{i=1}^{r}\alpha_k$,and $\alpha_k > 0 \quad k = 1,\ldots,r$

The posterior distribution $p(\theta \mid D,\xi) = Dir(\theta \mid \alpha_1 + N_1,\ldots,\alpha_r + N_r)$. Techniques for assessing the beta distribution, including the methods of imagined future data and equivalent samples, can also be used to assess Dirichlet distributions. Given this conjugate prior and data set D, the probability distribution for the next observation is given by

$$p(X_{N+1} = x^k \mid D,\xi) = \int \theta_k Dir(\theta \mid \alpha_1 + N_1,\ldots,\alpha_r + N_r)d\theta = \frac{\alpha_k + N_k}{\alpha + N} \tag{6}$$

As we shall see, another important quantity in Bayesian analysis is the marginal likelihood or evidence p(D|ξ). In this case, we have

$$p(D \mid \xi) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)}.\prod_{k=1}^{r}\frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)} \tag{7}$$

Namely, in the classical approach, θ is fixed (albeit unknown), and we imagine all data sets of size N that may be generated by sampling from the binomial distribution determined by θ. Each data set D will occur with some probability p(D|θ) and will produce an estimate θ*(D). To evaluate an estimator, we compute the expectation and variance of the estimate with respect to all such data sets:

$$Var_{p(D|\theta)}(\theta^*) = \sum_{D} p(D \mid \theta)(\theta^*(D) - E_{p(D|\theta)}(\theta^*))^2 \tag{8}$$

We then choose an estimator that somehowbalances the bias $(\theta - E_{p(D|\theta)}(\theta^*))$ and variance of these estimates over the possible values for θ. Finally, we apply this estimator to the data set that we actually observe. A commonly-used estimator is the maximum-likelihood (ML) estimator, which selects the value of θ that maximizes the likelihood p(D|θ). For binomial sampling, we have

$$\theta^*_{ML}(D) = \frac{N_k}{\sum_{k=1}^{r} N_k}$$

In contrast, in the Bayesian approach, D is fixed, and we imagine all possible values of θ from which this data set could have been generated. Given θ, the "estimate" of the physical probability of heads is just θ itself. Nonetheless, we are uncertain about θ, and so our final estimate is the expectation of θ with respect to our posterior beliefs about its value:

$$E_{p(\theta|D,\xi)}(\theta) = \int \theta p(\theta \mid D, \xi) d\theta \tag{9}$$

The expectations in Eqs. (8) and (9) are different and, in many cases, lead to different "estimates". One way to frame this difference is to say that the classical and Bayesian approaches have different definitions for what it means to be a good estimator. Both solutions are "correct" in that they are self consistent. Unfortunately, both methods have their drawbacks, which has lead to endless debates about the merit of each approach. Consequently, in this paper, we concentrate on the Bayesian approach.

Bayesian networks

A Bayesian network for a set of variables X ={X1,…, Xn} consists of 1.a network structure S that encodes a set of conditional independence assertions about variables in X, and 2.a set P of local probability distributions associated with each variable. Together, these components define the joint probability distribution for X. The network structure S is a directed acyclic graph. The nodes in S are in one-to-one correspondence with the variables X. We use Xi to denote both the variable and its corresponding node, and Pai to denote the parents of node Xi in S as well as the variables corresponding to those parents. The lack of possible arcs in S encode conditional independencies. In particular, given structure S, the joint probability distribution for X is given by

$$p(x) = \prod_{i=1}^{n} p(x_i \mid pa_i) \tag{10}$$

When building Bayesian networks from prior knowledge alone, the probabilities will be Bayesian. When learning these networks from data, the probabilities will be physical. To illustrate the process of building a Bayesian network, consider the problem of detecting credit-card fraud. We begin by determining the variables to model. One possible choice

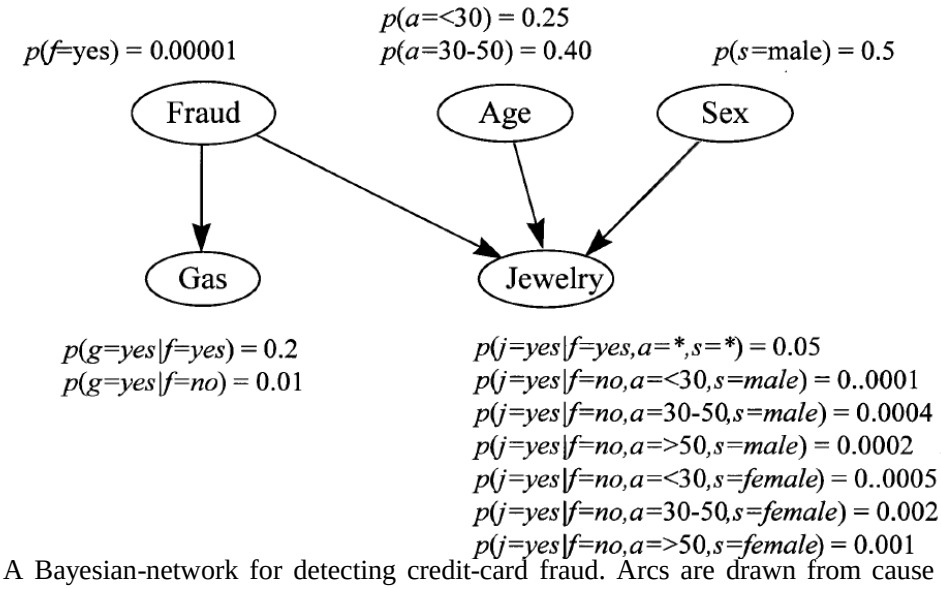

Figure 1. A Bayesian-network for detecting credit-card fraud. Arcs are drawn from cause to effect. The local probability distribution(s) associated with a node are shown adjacent to the node. An asterisk is a shorthand for "any state".

Of variables for our problem is Fraud (F), Gas (G), Jewelry (J), Age (A), and Sex (S), Of course, in a realistic problem, we would include many more variables. Also, we could model the states of one or more of these variables at a finer level of detail. This initial task is not always straight forward. As part of this task we must 1.correctly identify the goals of modeling, 2.identify many possible observations that may be relevant to the problem, 3.determine what subset of those observations is worthwhile to model, and 4.organize the observations into variables having mutually exclusive and collectively exhaustive states. Difficulties here are not unique to modeling with Bayesian networks, but rather are common to most approaches[7].

In the next phase of Bayesian-network construction, we build a directed acyclic graph that encodes assertions of conditional independence. One approach for doing so is based on the following observations. From the chain rule of probability, we have

$$p(x) = \prod_{i=1}^{n} p(x_i \mid \pi_i) \qquad (11)$$

Comparing Eqs.(11),we see that the variables sets(Π1,…,Πn) correspond to the Bayesian-network parents (Pa1,…,Pan), which in turn fully specify the arcs in the network structure S.

This approach has a serious drawback. If we choose the variable order carelessly, the resulting network structure may fail to reveal many conditional independencies among the variables [8]. In the worst case, we have to explore n! variable orderings to find the best one. Fortunately, there is another technique for constructing Bayesian networks that does not require an ordering. The approach is based on two observations [9]: 1.people can often readily assert causal relationships among variables, and 2.causal relationships typically correspond to assertions of conditional dependence. Example distributions are shown in fig .1.

As the amount of such data grows, it becomes harder and harder for humans to manage it and glean useful and reliable knowledge from it. This points to a growing need for automated tools to help us with this network engineering problem[10]. Please enjoy these interesting and significant contributions to the field of data mining for network engineering, and feel free to send your comments to me.

References

[1] Bielza, C., Li, G., Larranaga, P. (2011). Multi-dimensional classification with Bayesian networks. International Journal of Approximate Reasoning, 52(6), 705–727.

[2] Mihaljevic, B., Bielza, C., Larrañaga, P. (2013). bayesslass: an R package for learning Bayesian network classifiers. In Proceedings of useR!—the R user conference (p. 53).

[3] Morales, D., Vives-Gilabert, Y., Gómez-Ansón, B., Bengoetxea, E., Larrañaga, P., Bielza, C., Pagonabarraga, J., Kulisevsky, J., Corcuera-Solano, I., Delfino, M. (2013). Predicting dementia development in Parkinson's disease using Bayesian network classifiers. *Psychiatry Research: NeuroImaging*, *213*, 92–98.

[4] Balasundaram, B., Butenko, S., & Hicks, I. V. (2011). Clique relaxations in social network analysis: the maximum *k*-plex problem. *Operations Research*, *59*, *133–142*.

[5] Pattillo, J., Youssef, N., & Butenko, S. (2013). On clique relaxation models in network analysis. *European Journal of Operational Research*, *226*, *9–18*.

[6] Elkan, C. (2001). Magical thinking in data mining: Lessons from CoIL Challenge 2000. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , *426–431*.

[7] G. Provan, P. Langley, and P. Smyth. Bayesian Network Classifiers. *Machine Learning, 29, 131–163 (1997)*

[8] Agrawal D, Aggawal CC (2001) On the design and quantification of privacy preserving data-mining algorothms. In: Proceedings of the 20th ACM SIMOD symposium on principles of database systems. Santa Barbara, *247–255*

[9] Agrawal R, Srikant R (2000) Privacy-preserving data mining. In: Proceeding of the ACM SIGMOD conference on management of data. ACM Press, Dallas, TX, *439–450*

[10] Park BH, Kargupta H (2003) In: Nong Ye (ed) The handbook of data mining. *Lawrence Erlbaum Associates Inc Publishers, 341*