

Chapter 2

Probabilistic Graphical Models

Abstract This chapter introduces probabilistic graphical models as a statistical–structural pattern recognition paradigm. Many pattern recognition problems can be posed as labeling problems to which the solution is a set of linguistic labels assigned to extracted features from speech signals, image pixels, and image regions. Graphical models use Markov properties to measure a local probability on the labels within the neighborhood system. The Bayesian decision theory guarantees the best labeling configuration according to the *maximum a posteriori* criterion.

2.1 The Labeling Problem

As mentioned in Chap. 1, many pattern recognition problems can be posed as labeling problems to which the solution is a set of linguistic labels assigned to extracted features from speech signals, image pixels, and image regions. For instance, in speech recognition, we may have labels representing phonemes, and such a label set for the word “cat” would have labels for $/k/$, $/a/$, and $/t/$; in stroke segmentation of Chinese characters, we may have labels representing directions, and each character pixel may be associated with one of labels for horizontal, left-diagonal, vertical, and right-diagonal strokes; in Chinese character recognition, we may have labels representing stroke segments, which constitute different character structures; in topic modeling, we may have labels representing topics, which are the basic thematic components for a document. The labeling problem is shown in Fig. 2.1.

We specify a labeling problem in terms of a set of sites, $1 \leq i \leq I$, and a set of linguistic labels, $1 \leq j \leq J$; the j th label at any site i is denoted by $s_i = j$. A particular labeling configuration for the whole sites is denoted by $\mathcal{S} = \{s_1, s_2, \dots, s_I\}$. Note that the system can have the same label j at different sites, and not every label needs to be assigned to sites.

The sites may be successive times, image pixels, or image regions. We call them “regular” sites if they have the natural ordering, as for instance successive times form a one-dimensional sequence, in which sites $i - 1$ and $i + 1$ are before and behind site i ; image pixels form a lattice, where sites $(i', i - 1)$ and $(i', i + 1)$ are on the left and right side of site (i', i) . On the other hand, “irregular” sites have no natural

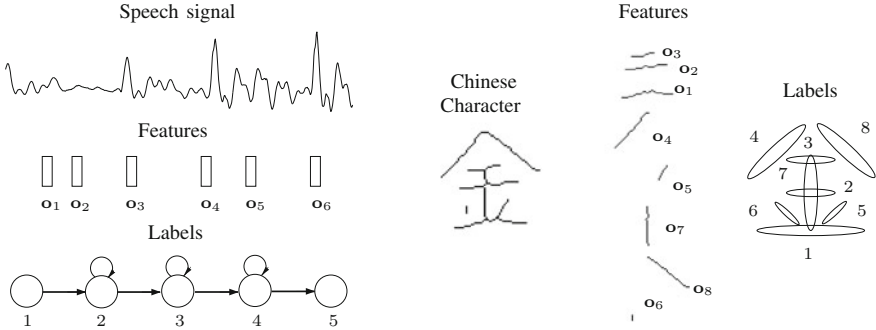


Fig. 2.1 Many pattern recognition problems can be posed as labeling problems

ordering such as image regions. We can define the ordering of irregular sites when necessary.

The linguistic labels can reflect any relations, regularities, or structures inherent in sites. For instance, phonemes can be divided into three stationary parts— initial, central, and final parts— we may use three labels to represent three successive parts of phoneme data. Chinese characters can be decomposed into blob-level regions (stroke segments), and these regions may be associated with linguistic labels representing character structures. Although labels can take continuous numerical values at each site such as image pixel intensities, we mainly focus on discrete linguistic labels in this book.

2.2 Markov Properties

Graphical models assume the label s_i is a random variable at site i , thus the labeling configuration \mathcal{S} is a stochastic process. Following the labeling problem, we put a probability measure, namely Markov properties, on the set of all possible labeling configurations,

$$P(\mathcal{S}) > 0, \forall \mathcal{S}, \quad (2.1)$$

$$P(s_i | \mathcal{S}_{\setminus i}) = P(s_i | \mathcal{N}_i), \quad (2.2)$$

where $\mathcal{S}_{\setminus i}$ are labels at all other sites except i , and \mathcal{N}_i are all labels at neighboring sites of i . This means that the probability of the label s_i is conditionally independent of all other labels except its neighboring labels. The neighborhood system \mathcal{N} plays an important role to reduce the global measure $P(s_i | \mathcal{S}_{\setminus i})$ to the local measure $P(s_i | \mathcal{N}_i)$, which may significantly simplify the computational complexity in practice. If we define the neighborhood system at two-dimensional sites, such as image pixels or regions, we call this graphical model the Markov random field (MRF) [1].

If we define the neighborhood system at one-dimensional sites, such as successive times, satisfying $\mathcal{N}_i = i - 1$, that is,

$$P(s_i | s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_I) = P(s_i | s_{i-1}), \quad (2.3)$$

we call this graphical model the first-order discrete-time Markov model.

Graphical models represent dependencies among random variables by a graph. Nodes in the graph are equivalent to sites, and edges between nodes imply some relationships between them. Obviously, the MRF is an undirected graphical model with Markov properties, and the first-order discrete-time Markov model has Markov properties on a directed chain graph. Markov properties simplify the global constraints $P(s_i | \mathcal{S}_{\setminus \{i\}})$ for all sites to the local constraints $P(s_i | \mathcal{N}_i)$ for neighboring sites, which greatly reduces the search space to find the best labeling configuration. Moreover, such a simplification is reasonable for piece-wise stationary data in terms of time and space.

Causality is an important property of the neighborhood system \mathcal{N} . If \mathcal{N}_i is defined by site i 's *previous* sites, it is strictly causal: The probabilities depend only on previous sites. Because regular sites have a natural ordering, it is easy to define the concept “previous.” For instance, the \mathcal{N} of the first-order discrete Markov model is causal, because time $i - 1$ always happens before time i ; the MRF may also have the causal neighborhood system: If we scan the image pixels from left to right and from up to down, then the previous sites of (i', i) are $(i' - 1, i)$ and $(i', i - 1)$. On the other hand, we often define a non-causal neighborhood system at irregular sites, such as image regions, because the concept “previous” at irregular sites is inconsistent in different conditions.

The causal neighborhood systems reduces the search space from “previous” and “following” to only “previous”. Such “cause-and-effect” relations are amenable to dynamic programming leading to high computational efficiency. For instance, we can search the best labeling configuration, $\mathcal{S}^* = \{s_1^*, \dots, s_I^*\}$, for a time sequence as follows: (1) First find the best label s_1^* , and then (2) find the best label s_2^* based on the best label s_1^* . We repeat this process until s_I^* .

2.3 The Bayesian Decision Theory

In practice, we have to bridge the gap between labels and data so as to characterize any relations, regularities, or structures inherent in some source of data. After assigning a label s_i to site i , we assume that s_i simultaneously generates an observation \mathbf{o}_i , which may be symbols, feature vectors, or image pixel values. At each labeling configuration, we will have an observation set, $\mathbf{O} = \{\mathbf{o}_1, \dots, \mathbf{o}_I\}$, at all sites, where \mathbf{O} belongs to the observation space Ω^I .

In the labeling problem, the task of pattern recognition is equivalent to finding a model λ that can provide the best labeling configuration, $\mathcal{S}^* = \{s_1^*, \dots, s_I^*\}$, to interpret observations, $\mathbf{O} = \{\mathbf{o}_1, \dots, \mathbf{o}_I\}$. The Bayesian decision theory is a fundamental

statistical approach to the problem of pattern recognition. Given a set of observations \mathbf{O} and class models λ_ω , we classify \mathbf{O} to the class ω^* by maximizing the posterior probability,

$$\omega^* = \arg \max_{\omega} \{P(\omega|\mathbf{O}, \lambda_\omega)\}. \quad (2.4)$$

By employing the Bayes formula,

$$P(\omega|\mathbf{O}, \lambda_\omega) = \frac{P(\mathbf{O}|\lambda_\omega)P(\omega)}{P(\mathbf{O})}, \quad (2.5)$$

where $P(\mathbf{O}|\lambda_\omega)$ is the likelihood of λ_ω given \mathbf{O} , and $P(\mathbf{O})$ is a constant normalization factor for all classes ω . For simplicity, we often assume equal prior class probability $P(\omega)$, so that the posterior probability is proportional to the likelihood,

$$P(\omega|\mathbf{O}, \lambda_\omega) \propto P(\mathbf{O}|\lambda_\omega). \quad (2.6)$$

Then Eq. (2.4) becomes

$$\omega^* = \arg \max_{\omega} P(\mathbf{O}|\lambda_\omega), \quad (2.7)$$

which is the *maximum-likelihood* (ML) criterion.

In graphical models, the class model λ_ω is a set of parameters λ and labels \mathcal{S} . We assign a labeling configuration \mathcal{S} to the observations \mathbf{O} with a joint probability $P(\mathcal{S}|\lambda, \mathbf{O})$. Thus, the $P(\mathbf{O}|\lambda)$ in Eq. (2.7) is computed by summing over all possible configurations,

$$P(\mathbf{O}|\lambda) = \sum_{\mathcal{S}} P(\mathcal{S}, \mathbf{O}|\lambda). \quad (2.8)$$

Because the direct computation of Eq. (2.8) is usually an intractable combinatorial problem, we can approximate the likelihood $P(\mathbf{O}|\lambda)$ by the most likely labeling configuration, $\mathcal{S}^* = \{s_1^*, \dots, s_I^*\}$, i.e.,

$$P(\mathbf{O}|\lambda) \approx P(\mathcal{S}^*, \mathbf{O}|\lambda). \quad (2.9)$$

Again the optimal labeling configuration \mathcal{S}^* for the observations \mathbf{O} can be obtained by maximizing the following posterior probability,

$$\mathcal{S}^* = \arg \max_{\mathcal{S}} P(\mathcal{S}|\mathbf{O}, \lambda). \quad (2.10)$$

From the Bayes formula,

$$P(\mathcal{S}|\mathbf{O}, \lambda) = \frac{P(\mathcal{S}, \mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} = \frac{p(\mathbf{O}|\mathcal{S}, \lambda)P(\mathcal{S}|\lambda)}{P(\mathbf{O}|\lambda)}, \quad (2.11)$$

where $P(\mathbf{O}|\lambda)$ is a constant normalization factor for all configurations \mathcal{S} . Thus, we obtain

$$P(\mathcal{S}|\mathbf{O}, \lambda) \propto p(\mathbf{O}|\mathcal{S}, \lambda)P(\mathcal{S}|\lambda), \quad (2.12)$$

where $p(\mathbf{O}|\mathcal{S}, \lambda)$ is the likelihood function for \mathcal{S} given \mathbf{O} , and $P(\mathcal{S}|\lambda)$ is the prior probability of \mathcal{S} . Therefore, Eq. (2.10) can be rewritten as

$$\mathcal{S}^* = \arg \max_{\mathcal{S}} p(\mathbf{O}|\mathcal{S}, \lambda)P(\mathcal{S}|\lambda). \quad (2.13)$$

When we have the knowledge about the data distribution but no appreciable prior information about the pattern, we may use the ML criterion to estimate the best labeling configuration. When the situation is the opposite, that is, when we have only prior information, we may use the principle of maximum entropy to find the least biased model that encodes the prior information. With both sources of information available, the best labeling configuration we can get is based on the Bayesian decision theory.

The Bayesian decision theory can be interpreted as a weighting mechanism that weighs the likelihood and prior distributions, and combines them to form the posterior. If these two distributions overlap significantly, this mathematical combination produces a desirable result. Otherwise, it may be possible that the posterior will fall into the region unsupported by either the likelihood or the prior. Therefore, in real applications, we have to balance the likelihood and prior to achieve a desirable labeling configuration.

Note that the likelihood function $p(\mathbf{O}|\mathcal{S}, \lambda)$ is not a probability but a subjective function, which enables us to assign relative weights to different configurations \mathcal{S} given \mathbf{O} . On the other hand, the prior probability $P(\mathcal{S}|\lambda)$ is the source of information that exists *prior* to test data in the form of expert judgement and other historical information (training data). For instance, we may specify some structural information in the labeling space based on the knowledge of the problem domain, or just obtain this information from training data automatically. The Bayesian decision rule provides a convenient method to combine two different sources of information, i.e., the data and the prior. It is advantageous to combine distribution functions from different information sources in that the uncertainty in the posterior distribution is reduced when the information in the likelihood and prior distributions are consistent with each other. In other words, the integral information from the data and the prior distribution may have less uncertainty because the data and the prior are from two different information sources that may support each other. From the regularization point of view, the combination of the likelihood and the prior may convert a mathematically ill-posed recognition problem into a well-posed one.

2.3.1 Descriptive and Generative Models

We further investigate the graphical models within Bayesian framework in the view of descriptive and generative models.

Descriptive methods construct the model for a pattern by imposing statistical constraints on features extracted from patterns. Linguistic labels can be viewed as high-level features extracted from patterns. Therefore, graphical models are descriptive models, as for instance the first-order discrete Markov model imposes statistical constraints for labels at successive times; the MRF imposes local statistical constraints of labeling configuration at neighboring image pixels or regions. Descriptive models specify the structural constraints $P(s_i | \mathcal{S}_{\setminus i})$ in the labeling space either specified by prior knowledge or learned through training samples. We may view descriptive constraints as the necessary condition for the pattern class, in which all samples in the pattern class must satisfy these constraints (with high probability), while samples from other classes may also satisfy such constraints. Figure 2.2 shows the relationship between descriptive models and samples, where the circle of descriptive models contains all samples as the necessary condition. Obviously, such a necessary condition is not enough to model the difference among pattern classes, because different pattern classes may probably share the same structure or substructure in the labeling space.

In contrast to descriptive models, generative models are able to randomly generate observed data, typically given some hidden variables at sites. Linguistic labels are a kind of hidden variables, and after randomly assigned to all sites, they simultaneously generate observations \mathbf{O} . The hidden variables employed to generate observations usually follow very simple models, such as Gaussian mixture models (GMMs). Furthermore, existing generative models appear to suffer from an oversimplified assumption that the observations are independent and identical distributed (i.i.d.)

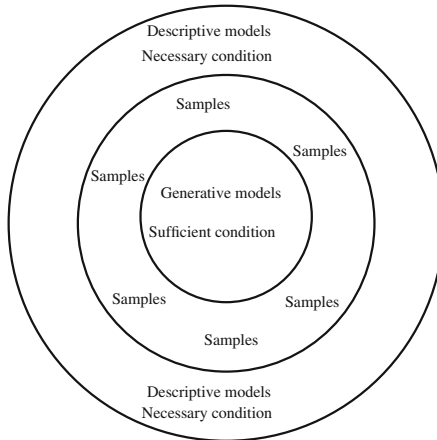


Fig. 2.2 Descriptive models are the necessary condition and generative models are the sufficient condition for samples from a specific pattern class

for all sites. So the joint probability of observations can be written as a product of probabilities of individual observations,

$$p(\mathbf{O}|\mathcal{S}) = \prod_{i=1}^I p(\mathbf{o}_i | s_i = j), \quad 1 \leq j \leq J. \quad (2.14)$$

As a result, generative models are not sufficient enough to model realistic patterns. We may view generative models as the sufficient condition for the pattern class, in which the model generates all samples belonging to the pattern class (with high probability), whereas some samples from the pattern class may not be generated by the model (with low probability). Figure 2.2 shows the relationship between generative models and samples, where generative models are subsets of samples as the sufficient condition.

Bayesian decision rule integrates both descriptive and generative models to provide the sufficient and necessary conditions for the pattern class in a hierarchical system. The bottom level of the system is generative in nature, because the observations are generated by hidden variables such as labels. The top level of the system is descriptive in nature, because it governs the relationships among random hidden variables probabilistically. For instance, in speech recognition, we use labels (GMMs) to generate speech feature vectors at each time i , and use transition probability $a_{jj'}$ to control the jump from label j to j' ; in Chinese character recognition, we use labels (GMMs) to generate stroke segments, and use prior clique potentials to encourage or penalize different local labeling configurations.

2.3.2 Statistical–Structural Pattern Recognition

The integration of descriptive and generative models also combines both structural and statistical information of the pattern. Descriptive models $P(\mathcal{S})$ can describe the high-level structure of linguistic labels, while generative models $p(\mathbf{O}|\mathcal{S})$ can describe low-level statistical uncertainty of observations. Therefore, graphical models with the Bayesian decision rule provide a theoretically well-founded framework to represent both structural and statistical information existing universally in pattern recognition problems. We call this paradigm the statistical–structural pattern recognition.

2.4 Summary

Labeling problems have been proposed to solve problems of computer vision and image analysis in [2]. In this chapter, we extend the same concept to more general problems of pattern recognition. The Bayesian theory has long been the dominant classification methods in pattern recognition, because it rests on an axiomatic

foundation that is guaranteed to have quantitative coherence; some other classification methods may not [3]. Further study of the role of the Bayesian theory in fuzzy logic can be found in [4]. The formulation of Bayesian framework for labeling problems is actually the compound Bayes decision problem by the use of context information [3], in which the states of nature are equivalent to the labels here. Markov properties simplify the interdependence of labels by assuming that the labels are only dependent on their neighbors, which avoids the computation of $P(\mathcal{S})$ for all J^I possible values of labeling configuration. Therefore, graphical models have been widely applied to problems of pattern recognition [3], such as speech recognition, handwriting recognition, gesture recognition, face recognition, human motion recognition, and DNA sequence recognition.

The relationship between descriptive models and generative models has been discussed in [5, 6]. We use this idea to justify the modeling ability of graphical models for labeling problems. Graphical models can integrate both descriptive and generative models so that they satisfy the sufficient and necessary conditions to model samples from pattern classes. Murphy and Smyth considered Markov models are special cases of graphical models [7–9], in which HMMs and MRFs are acyclic directed graphs and undirected graphs with Markov properties, respectively. The graph represents the structure of random variables, so graphical models can represent structural patterns statistically [10]. Taking higher order of statistical dependencies between labels into account, graphical models can indirectly reflect statistical dependencies between observations, despite the conditionally independent assumption upon observations in terms of the labels.

References

1. Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition. Springer, New York (2005)
2. Li, S.Z.: Markov Random Field Modeling in Image Analysis. Springer-Verlag, New York (2001)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
4. Ross, T.J., Booker, J.M., Parkinson, W.J.: Fuzzy Logic and Probability Applications: Bridging the Gap, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2002)
5. Zhu, S.C.: Statistical modeling and conceptualization of visual patterns. *IEEE Trans. Pattern Anal. Machine Intell.* **25**(6), 691–712 (2003)
6. Guo, C.E., Zhu, S.C., Wu, Y.N.: Modeling visual patterns by integrating descriptive and generative methods. *International Journal of Computer Vision* **53**(1), 5–29 (2003)
7. Murphy, K.: A brief introduction to graphical models and Bayesian networks (1998). <http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html>
8. Smyth, P.: Belief networks, hidden Markov models, and Markov random fields: a unifying view. *Pattern Recognition Letters* **18**(11–13), 1261–1268 (1997)
9. Lauritzen, S.L.: Graphical Models. Clarendon Press, Oxford (1996)
10. Zeng, J., Liu, Z.Q.: Markov random fields-based statistical character structure modeling for Chinese character recognition. *IEEE Trans. Pattern Anal. Machine Intell.* **30**(5), 767–780 (2008)

Type-2 Fuzzy Graphical Models for Pattern Recognition

Zeng, J.; Liu, Z.-Q.

2015, XIII, 201 p. 112 illus., Hardcover

ISBN: 978-3-662-44689-8