Markov Random Fields with Applications to M-reps Models

Conglin Lu

Medical Image Display and Analysis Group

University of North Carolina, Chapel Hill

Markov Random Fields with Applications to M-reps Models

Outline:

- ❑ Background;
- ❑ Definition and properties of MRF;
- ❑ Computation;
- ❑ MRF m-reps models.

Markov Random Fields

Model a large collection of random variables with complex dependency relationships among them.

Markov Random Fields

- A model based approach;
- Has been applied to a variety of problems:

  - Speech recognition
  - Natural language processing
  - Coding
  - Image analysis
  - Neural networks
  - Artificial intelligence

- Usually used within the Bayesian framework.

The Bayesian Paradigm

X = space of the unknown variables,

e.g. labels;

Y = space of data (observations),

e.g. intensity values;

Given an observation $y \in Y$, want to make inference about $x \in X$.

The Bayesian Paradigm

Prior P_X : probability distribution on X;

Likelihood $P_{Y|X}$: conditional distribution of Y given X;

Statistical inference is based on the *posterior* distribution $P_{X|Y} \propto P_X \bullet P_{Y|X}$.

The Prior Distribution

- Describes our assumption or knowledge about the model;
- X is usually a high dimensional space. P_X describes the joint distribution of a large number of random variables;
- How do we define P_X?

Markov Random Fields with Applications to M-reps Models

Outline:

- ✓ Background;
- □ Definition and properties of MRF;
- □ Computation;
- □ MRF m-reps models.

Assumptions

- $X = \{X_s\}_{s\in\mathcal{S}}$, where each X_s is a random variable; $\mathcal{S}$ is an index set and is finite;
- There is a common state space $\mathcal{R}$: $X_s \in \mathcal{R}$ for all $s \in \mathcal{S}$; $|\mathcal{R}|$ is finite;
- Let $\Omega = \{\omega=(x_{s_1}, \ldots, x_{s_N}): x_{s_i} \in \mathcal{R}, 1 \leq i \leq N\}$ be the set of all possible *configurations*.

Dependency Graph

A simple undirected graph $\mathcal{G} = (\mathcal{S}, \mathcal{N})$:

- $\mathcal{S}$ is the set of sites (vertices);
- $\mathcal{N}=\{\mathcal{N}_s\}_{s\in\mathcal{S}}$ is the neighborhood structure (the set of edges). The neighbors of s are those sites that are connected to s by an edge;
- Let $\mathcal{C}$ denote the set of cliques - completely connected subgraphs of $\mathcal{G}$, including singletons.

Markov Random Field: Definition

P is an Markov random field on Ω with respect to $\mathcal{G} = (\mathcal{S}, \mathcal{N})$ if

(1) $P(X=\omega) > 0$ for all $\omega \in \Omega$;

(2) $P(X_s=x_s \mid X_r=x_r, r \neq s)$

$= P(X_s=x_s \mid X_r=x_r, r \in \mathcal{N}_s)$

(local characteristics)

The local characteristics uniquely determines a joint distribution.

Examples of MRF: Nearest Neighbor Systems

- 1st order Markov chain $\{X_0, X_1, \dots, X_n, \dots\}$:

  $P(X_{n+1}=x_{n+1} \mid X_n=x_n, X_{n-1}=x_{n-1}, \dots, X_0=x_0)$

  $= P(X_{n+1}=x_{n+1} \mid X_n=x_n)$

- 4-neighbor lattice system:

  $P(X_{i,j} \mid \text{all other random variables}) =$

  $P(X_{i,j} \mid X_{i-1,,j}, X_{i+1,j}, X_{i,,j-1}, X_{i,,j+1})$

Gibbs Field

P is Gibbs on Ω with respect to $\mathcal{G} = (\mathcal{S}, \mathcal{N})$ if

$$P(\omega) = 1/Z \cdot \exp\{-H(\omega) / T\},$$

where

- Z is a normalizing constant (*partition function*);
- H is the *energy*. $H(\omega) = \sum_{C \in \mathcal{C}} U_C(\omega)$. $\mathcal{C}$ is the set of cliques for $\mathcal{G}$. $\{U_C \geq 0\}$ are called the *potentials*;
- $U_C(\omega)$ depends only on those x_s of ω for which $s \in C$;
- T is a parameter (*temperature*).

The Hammersley-Clifford Theorem

P is an MRF with respect to $\mathcal{G}$ if and only if P is a Gibbs distribution with respect to $\mathcal{G}$.

Advantage of Using the Gibbs Form

- The Gibbs form explicitly specifies the joint distribution;
- Local characteristics (conditional probabilities) can be easily formulated from the Gibbs form;
- The potentials can be learned from training data (see later slides) .

Examples: Nearest Neighbor Systems (cont.)

- 1-D :

  $$H(\{x_i\}) = \sum U_i(x_i) + \sum U_{(i,i+1)}(x_i, x_{i+1})$$

- 2-D :

  The most general form of the energy is

  $$H(\{x_{i,j}\}) = \sum U_{\{(i,j)\}}(x_{i,j})$$
  $$+ \sum U_{\{(i,j),\,(i+1,j)\}}(x_{i,j}, x_{i+1,j})$$
  $$+ \sum U_{\{(i,j),\,(i,j+1)\}}(x_{i,j}, x_{i,j+1})$$

Important Properties of MRF

- Markov property:

  Let A, B, C $\subset$ S. If every path from a$\in$A to c$\in$C meets some b$\in$B, then X_A and X_C are conditionally independent given X_B.

  Can still model complicated dependencies!

- Maximum entropy property:

  The family $P_\lambda(\omega) = 1/Z_\lambda \exp\{- \Sigma_c \lambda_c U_c(\omega)\}$ are the maximum entropy models with fixed values for $E(U_c(\omega)) = U^*_c$ (average energy).

Learning by the ME Principle

- Choose a set of (local) features;
- Obtain empirical distribution of the features from training set;
- Learn the potentials by the ME principle.
- Example: ME distribution with specified mean and variance yields a Gaussian distribution.

Markov Random Fields with Applications to M-reps Models

Outline:

✓ Background;

✓ Definition and properties of MRF;

☐ Computation;

☐ MRF m-reps models.

Computation Methods

- Dynamic programming
  - the basic idea behind a lot of different algorithms, e.g. forward-backward, parsing, Viterbi, sum-product, belief propagation, etc.;
  - relatively fast;
  - does not work for all MRF’s.
- Stochastic relaxation

General Computation Problems

a) Sample from a Gibbs distribution;
b) Find minimum energy;
c) Compute expected values;
d) Test model and estimate parameters.

Among them, a) is the most basic problem.

Direct sampling from a Gibbs field $P(x) = Z^{-1} \exp(-H(x))$, $x \in X$, is usually not feasible because

- the underlying space X is huge;
- the partition function Z is intractable.

Stochastic Sampling Algorithms

Design a Markov chain with state space Ω whose equilibrium distribution is the desired Gibbs distribution.

Examples:

- Metropolis -Hastings algorithms: based on having "elementary" Markov chains;
- Gibbs sampler: based on using local characteristics.

Temperature in Gibbs Distribution

Any Gibbs field P can be put in a family $\{P_T\}$ with parameter T = temperature:

$$P_T(x) = 1/Z_T \, P(x)^{1/T}$$

$$= 1/Z_T \cdot \exp\{-E(x)/T\},$$

- as $T \to \infty$, $P_T \to$ uniform distribution;
- as $T \to 0$, $P_T \to \delta_{\text{mode}(P)}$.

Simulated Annealing

- Goal: find the global minimum energy (*ground state*), e.g. MAP estimates.
- Algorithm:
  - choose a cooling scheme $T_1 > T_2 > \ldots \rightarrow 0$;
  - generate a Markov chain $\{X^{(n)}\}$ on Ω where $X^{(n)} \rightarrow X^{(n+1)}$ is a transition by P^{Tn} ; the transition probabilities are specified by the Metropolis/Gibbs sampler;
  - If one cools at a *very slow* pace, then $X^{(n)}$ converges in probability to the mode of P.

Simulated Annealing (cont.)

- Advantages:
  - – guaranteed to find global minima (in principle), as opposed to greedy algorithms;
  - – works for any Gibbs fields;
- Disadvantages:
  - – convergence is very slow;
  - – stopping rule is not clear;
  - – hard to analyze;

Markov Chain Monte Carlo

- Goal: compute $E_p(f)$ for a function f on Ω.
- Traditional Monte Carlo: sample uniformly from Ω and average w.r.t. P

$$E_P(f(X)) \approx \sum_k f(X_k)P(X_k) \Big/ \sum_k P(X_k)$$

- MCMC: sample from P and average uniformly

$$E_P(f(X)) \approx \sum_{k=1}^{K} f(X_k)/K$$

Summary of the Theory

- MRF provides a general framework for studying complex random systems;
- Computation is usually complicated;
- How can we do better?
  - data driven methods in computation;
  - better design of MRF, e.g. hierarchical MRF modes (HMF, HMRF, etc.);
  - other approximations, e.g. mean field, continuous stochastic processes.

Markov Random Fields with Applications to M-reps Models

Outline:

- ✓ Background;
- ✓ Definition and properties of MRF;
- ✓ Computation;
- □ MRF m-reps models.

M-reps Models

- Multiscale shape models. Each scale *k* is described by a set of primitives $\{\mathbf{z}^{k}_{i}\}$;
- Object intrinsic coordinates provide correspondences among object population;
- Can easily describe both *global* and *local* variations, as well as inter-object relations.

MRF M-reps Models

- The probability distribution on the shape space is given by $P(\{\mathbf{z}^k_i\})$;
- Markov assumption:

  $P(\mathbf{z}^k_i \mid \text{all other primitives at all scales} \leq k)$

  $= P(\mathbf{z}^k_i \mid \mathcal{N}(\mathbf{z}^k_i), \mathcal{P}(\mathbf{z}^k_i))$
- If $\mathbf{z}^k$ denotes scale k, then a multiscale MRF m-reps model can be written as a Markov chain

  $$P(\mathbf{z}^1, \ldots \mathbf{z}^n) = P(\mathbf{z}^1) \cdot P(\mathbf{z}^2 \mid \mathbf{z}^1) \cdots P(\mathbf{z}^n \mid \mathbf{z}^{n-1})$$

MRF M-reps Models

- By the H-C theorem, the model has “two-sided” Markov property, i.e. $P(\mathbf{z}^k | \text{ all other scales}) = P(\mathbf{z}^k | \mathbf{z}^{k-1}, \mathbf{z}^{k+1})$, or equivalently,

  $P(\mathbf{z}^k_i \,|\, \text{all other primitives at all scales})$

  $= P(\mathbf{z}^k_i \,|\, \mathcal{N}(\mathbf{z}^k_i), \mathcal{P}(\mathbf{z}^k_i), \mathcal{C}(\mathbf{z}^k_i))$
- Use residues (differences) as features;
- The basic problem is how to specify the conditional probabilities $P_k = P(\mathbf{z}^k | \mathbf{z}^{k-1})$

The Boundary Level: MRF Model

- Primitives: $z_i = \tau_i$, the (normalized) displacement along the normal direction at point i;
- Neighborhood structure: nearest 4-neighbors;
- The Gibbs distribution thus involves potentials of the form $A_i(\tau_i)$ and $B_{ij}(\tau_i, \tau_j)$, where i and j are 4-neighbors.

The Boundary Level: MRF Model

- Further assumptions:
  - Potentials have the same function form;
  - Gaussian (quadratic potentials);
- The joint distribution of $\{\tau_i\}$ has density

$$P(\{\tau_i\}) = \frac{1}{Z}\exp\{-\frac{1}{2\sigma_1^2}\sum_i s_i\tau_i^2 - \frac{1}{2\sigma_2^2}\sum_{<i,j>} w_{ij}(\tau_i - \tau_j)^2\}$$

- σ_1, σ_2 are parameters; $\{s_i\}$and $\{w_{ij}\}$are fixed from the previous stage.

The Boundary Level: Conditional Distribution

$$P(\{\tau_i\}) = \frac{1}{Z}\exp\left\{-\frac{1}{2\sigma_1^2}\sum_i s_i\tau_i^2 - \frac{1}{2\sigma_2^2}\sum_{<i,j>} w_{ij}(\tau_i - \tau_j)^2\right\}$$

The log of the conditional probability density of τ_i is essentially

$$-\frac{s_i\tau_i^2}{2\sigma_1^2} - \frac{\sum_{j\in\mathcal{N}(i)} w_{ij}}{2\sigma_2^2}\left(\tau_i - \sum_{j\in\mathcal{N}(i)} \frac{w_{ij}}{\sum_{j\in\mathcal{N}(i)} w_{ij}}\tau_j\right)^2$$

<u>Interpretation</u>: penalizes large τ_i and large deviation from “predicted τ” by neighbors.

The Boundary Level: Prior Model Learning

- The parameters σ_1, σ_2 can be learned from training data, using maximum likelihood estimates or other criteria;
- Other choices of model:
  - – position-dependent parameters;
  - – non-Gaussian models, maximum entropy learning.

The Atom Level

- Primitive: $z_i = \mathbf{A}_i = (\mathbf{x}_i, \mathbf{R}_i, r_i)$, describing position $\mathbf{x}$, local frame $\mathcal{F}$, and radius r of atom i. $z_i \in \boldsymbol{R}^3 \times SO(3) \times \boldsymbol{R}^+$;
- With 4-neighbor structure, the Gibbs distribution contains potentials of the form $f_i\,(\mathbf{A}_i)$ and $g_{ij}(\mathbf{A}_i\,, \mathbf{A}_j\,)$ for neighboring atoms;
- Need a metric to describe difference between atoms …

Atom Distance

Define a metric on atoms (or $\boldsymbol{R}^3 \times \mathrm{SO}(3) \times \boldsymbol{R}^+$) by

$$d(\mathbf{A}_i, \mathbf{A}_j) = \sqrt{\alpha_E d_E^2(\mathbf{x}_i, \mathbf{x}_j) + \alpha_R d_R^2(\mathbf{R}_i, \mathbf{R}_j) + \alpha_r d_r^2(r_i, r_j)}$$

where

- – d_E is Euclidean distance in $\boldsymbol{R}^3$;
- – d_R is the Riemannian distance in SO(3);
- – d_r is the log-distance in $\boldsymbol{R}^+$: $d(r_1,r_2) = |\log(r_1/r_2)|$;
- – α_E, α_R, α_r are appropriate weights.

The Atom Level: MRF Model

- Let $\Delta\mathbf{A}_i$ denote the “difference” between $\mathbf{A}_i$ and $\mathbf{A'}_i$, where $\mathbf{A'}_i$ is the corresponding atom at the previous scale. In other words,

$$\Delta\mathbf{A}_i = (\Delta\mathbf{x}_i, \Delta\mathbf{R}_i, \Delta r_i\,) = \big((\mathbf{x}_i - \mathbf{x'}_{i)}) \,/\, r'_i,\ (\mathbf{R'}_i)^{-1}\mathbf{R}_i,\ r_i \,/\, r'_i\big)\ .$$

- Prior model (quadratic potentials):

$$P(\{\mathbf{A}_i\}\,|\,\{\mathbf{A}_i^{'}\}) = \frac{1}{Z}\exp\Big\{-\sum_i \frac{s_i}{2\sigma_i^2} d^2(\mathbf{A}_i, \mathbf{A}_i^{'})$$

$$-\sum_{<i,j>} \frac{w_{ij}}{2\sigma_{ij}^2} d^2(\Delta\mathbf{A}_i, \Delta\mathbf{A}_j)\Big\}$$

The Atom Level: Conditional Distribution

$$P(\mathbf{A}_i) \propto \exp\Big\{-\frac{S_i}{2\sigma_i^2}[\alpha_E d_E^2(\Delta\mathbf{x}_i, 0) + \alpha_R d_R^2(\Delta\mathbf{R}_i, \mathbf{I}) + \alpha_r d_r^2(\Delta r_i, 1)]$$

$$-\sum_{j\in\mathcal{N}(i)} \frac{w_{ij}}{2\sigma_{ij}^2}[\alpha_E d_E^2(\Delta\mathbf{x}_i, \Delta\mathbf{x}_j) + \alpha_R d_R^2(\Delta\mathbf{R}_i, \Delta\mathbf{R}_j) + \alpha_r d_r^2(\Delta r_i, \Delta r_j)]\Big\}$$

- σ_i , σ_{ij} are trainable parameters of the model;
- The density is with respect to the Haar measure on the product space $\boldsymbol{R}^3\times SO(3)\times\boldsymbol{R}^+$;
- Interpretation: penalties on being away from “parent atom” and “neighbor mean”;

MRF M-reps Models

- Similar MRF models can be designed for all other scale levels, using appropriate parent and neighbor terms;
- The full joint distribution is a probability measure on the shape space, with a relatively small number of parameters;
- The model is trainable (parametric vs. non-parametric).

References

- For Markov random fields:
  - Geman, Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", IEEE PAMI 6, No. 6, 1984;
- For deformable m-reps models:
  - Joshi, Pizer, et al, TMI 2002;
  - Pizer, Joshi, et al, IJCV 2002.