文章编号:1001 - 9081(2008)03 - 0699 - 04

基于多模板隐马尔可夫模型的文本信息抽取算法

胡宇舟[1],王 雷[2,3],顾学道[3]
(1. 天津大学 管理学院,天津 300072; 2. 清华大学 计算机科学与技术博士后流动站,北京 100084;
3. 深圳现代计算机有限公司 博士后科研工作站,广东 深圳 518057)
(guxd@mcm.com.cn)

摘 要:由于训练数据来源的多样化,难以通过学习得到最优的模型参数,因此提出了一种基于多模板隐马尔可夫模型的文本信息抽取算法。该算法首先利用文本排版格式和分隔符等信息,对文本进行分块;然后在分块的基础上,对训练数据进行聚类以形成多个形式的模板(多模板),并对多模板数据训练得到隐马尔可夫初始概率及转移概率参数;最后,用被训练的数据统一训练释放概率参数,结合初始概率、转移概率以及释放概率参数对文本信息进行抽取。实验结果表明,该算法在精确度和召回率指标上比简单隐马尔可夫模型具有更好的性能。

关键词:文本信息抽取;隐马尔可夫模型; 多模板;文本分块
中图分类号: TP18;TP391.1 **文献标志码**:A

Text information extraction algorithm based on multiple templates hidden Markov model

HU Yu-zhou[1], WANG Lei[2,3], GU Xue-dao[3]
(1. *School of Management, Tianjin University, Tianjin* 300072, *China*;
2. *Postdoctoral Program of Computer Science and Technology, Tsinghua University, Beijing* 100804, *China*;
3. *Postdoctoral Program, Shenzhen Modern Computer Manufacture, Shenzhen Guangdong* 518057, *China*)

Abstract: Since training data sources are varied and it is difficult to obtain optimal model parameters through learning, a text information extraction algorithm based on Hidden Markov Model (HMM) with multiple templates was proposed. Firstly the algorithm segmented texts by using the information of typesetting formats and list separators. Then multiple templates were formed with clustering the training data based on the segmentations, and the parameters of initial probability and transition probability for HMM were obtained to train data of the templates. Finally releasing probability parameters of the universal training were obtained with the trained data, and text information was extracted through combining the initial and transition probability parameters and the releasing probability parameters. Experimental results show that, the new algorithm has better performance in precision and recall than simple hidden Markov model.

Key words: text information extraction; Hidden Markov Model (HMM); multiple templates; text block

0 引言

自动文本信息抽取是文本信息处理的一个重要环节[1]。文本信息抽取(text information extraction)是指从文本中自动抽取相关的或特定类型的信息。目前文本信息抽取模型主要有三种:基于词典的抽取模型[2],基于规则的抽取模型[3]和基于隐马尔可夫模型(Hidden Markov Model,HMM)的抽取模型[4-8]。

利用 HMM 进行文本信息抽取是一种基于统计机器学习的信息抽取方法。HMM 易于建立,具有不需要大规模的词典集与规则集,适应性好和抽取精度较高等优点,因而得到研究者的关注。文献[4] 利用文本排版格式和分隔符等信息,对文本进行分块,在分块的基础上结合隐马尔可夫模型进行文本信息抽取;文献[5]使用一种“收缩”的技术改进 HMM 信息抽取模型概率的估计;文献[6]使用随机优化技术动态选择最适合的 HMM 模型结构进行信息抽取;文献[7]将自然语言处理中的短语结构分析技术应用到 HMM 文本信息抽取中来;文献[8]利用主动学习技术来减少训练 HMM 信息抽取模型时所需的标记数据;文献[9]通过聚类算法来改进 HMM 的文本详细抽取的精度。

网上不同来源的训练数据,在其排版、形式和表示风格上有时很不相同,甚至相差很大。举个例子来说,技术报告和一般的杂志期刊或学位论文的头部格式是很不同的,有时不同期刊杂志上的论文的格式也很不相同。常用的方法是将所有的训练数据混合起来训练隐马尔可夫模型参数,由于训练数据成分很不相同,因此很难通过统计学习技术得到最优的模型参数。针对训练数据来源的多样难以通过学习得到最优的模型参数的问题,本文提出了一种基于多模板隐马尔可夫模型的文本信息抽取算法。新算法首先利用文本排版格式和分隔符等信息,对文本进行分块;然后在分块的基础上,通过基于形式的聚类方法将训练数据聚成几个类,其中每个类代表一个模板;每个类(即每个模板)的数据被用来训练一个初始概率和一个转移概率矩阵,而所有的训练数据被用来训练一个统一的释放概率矩阵;最后,在进行文本信息抽取时,结合

收稿日期:2007 - 09 - 27;**修回日期**:2007 - 11 - 27。
基金项目:湖南省自然科学基金资助项目(03JJY3098);福建省青年科技人才创新项目(2005J051)。
作者简介:胡宇舟(1963 -),男,湖南长沙人,高级工程师,博士,主要研究方向:信息管理系统、人工智能; 王雷(1973 -),男,湖南长沙人,副教授,博士,主要研究方向:通信网络、数据挖掘; 顾学道(1939 -),男,上海人,教授,主要研究方向:通信网络、人工智能、软件工程。

每一个初始概率矩阵、每一个转移概率矩阵和统一的释放概率矩阵,使用 Viterbi 算法来找出最优的标记序列,这些最优的标记序列中概率最大的标记序列将被作为最终输出。实验结果表明,与简单隐马尔可夫模型相比,新算法在精确度和召回率指标上具有更好的性能。

1 基于 HMM 的文本信息抽取

1.1 HMM 模型

HMM 提供了一种基于训练数据的概率自动构造识别系统的技术。一个 HMM 包含两层:一个可观察层和一个隐藏层。可观察层是待识别的观察序列,隐藏层是一个马尔可夫过程,即是一个有限状态机,其中每个状态转移都带有转移概率[10]。

一个 HMM 应用在文本抽取时可以看成一个五元组 $\{S, V, A, B, \Pi\}$:

$$S = \{S_1, S_2, \cdots, S_N\}$$

$$V = \{V_1, V_2, \cdots, V_M\}$$

$$\boldsymbol{A} = \{a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i), 1 \leqslant i, j \leqslant N\}$$

$$\boldsymbol{B} = \{b_j(V_k) = P(V_k \text{ at } t \mid q_t = S_j), 1 \leqslant j \leqslant N, 1 \leqslant k \leqslant M\}$$

$$\Pi = \{\pi_i = P(q_1 = S_i), 1 \leqslant i \leqslant N\}$$

其中:S 是状态集,共有 N 个状态;V 是词汇集,共有 M 个可能的输出单词;$\boldsymbol{A}$ 是 $N \times N$ 的状态转移矩阵,a_{ij} 表示从状态 S_i 转换到状态 S_j 的概率;$\boldsymbol{B}$ 是 $N \times M$ 的释放概率矩阵,$b_j(V_k)$ 表示在状态 b_j 时释放单词 V_k 的概率;Π 是初始状态概率集合,π_i 是第 i 个状态作为初始状态的概率。

1.2 基于 HMM 的文本信息抽取模型

应用 HMM 模型,主要解决三个方面的关键问题,即:评估问题、学习问题和解码问题。对于文本信息抽取需要解决 HMM 模型中的学习问题和解码问题。进行信息抽取时,一般是先训练样本,对于已标记训练样本采用 ML(Maximum Likelihood)算法,或对于未标记训练样本采用 Baum-Welch 算法进行学习,得出 HMM 模型参数;然后采用 Viterbi 算法将待抽取的输入文本序列标记为最大概率的状态标签序列。一般地,基于 HMM 模型的文本信息抽取分为以下两大步骤[11]:

1)应用统计的方法从训练样本中得出 HMM 模型参数。采用 ML 算法,建立 HMM 模型。ML 算法主要以统计的方法得出 HMM 模型参数,由以下三个公式分别计算模型的初始状态概率 π_i、转移状态概率 a_{ij} 和状态释放概率 $b_j(V_k)$,即:

$$\pi_i = \frac{Init(i)}{\sum_{j=1}^{N} Init(j)}; \ 1 \leqslant i \leqslant N_S \tag{1}$$

其中,$Init(i)$ 是所有训练序列中,初始状态为 i 的序列个数。

$$a_{ij} = \frac{C_{i,j}}{\sum_{K=1}^{N} C_{i,k}}; \ 1 \leqslant i, j \leqslant N_S \tag{2}$$

其中,$C_{i,j}$ 是所有训练序列中,从状态 S_i 转换到状态 S_j 的次数。

$$b_j(V_k) = \frac{E_j(V_k)}{\sum_{i=1}^{M} E_j(V_i)}; \ 1 \leqslant j \leqslant N, \ 1 \leqslant k \leqslant M \tag{3}$$

其中,$E_j(V_k)$ 是所有训练序列中,状态 S_j 释放单词 V_k 的次数。

2)应用已建立好的 HMM 模型进行文本信息抽取。以文本观察序列 $O = O_1 O_2 \cdots O_T$ 作为模型输入,采用 Viterbi 算法,找出最大概率的状态标签序列,被标记为目标状态标签的观察文本即为信息抽取的内容。

2 基于文本分块的多模板 HMM 文本信息抽取

目前文献中使用的 HMM 信息抽取模型一般以单词作为基本抽取单位,考虑到文本排版格式和分隔符等信息的存在,实际上文本可以看作是由一些文本分块序列组成的。我们可以将一篇文本按照排版格式和分隔符等信息划分成文本分块序列,在信息抽取过程中采用结合文本分块信息的 HMM 模型。由于待抽取的信息实际上是由这些文本分块组成的,文本分块所包含的特征信息明显要比单个单词包含的特征信息多,以分块为基本单位更易于机器自动识别与抽取。文本分块思想是将论文头部划分为文本块序列,并保证这样分割后的分块足够小,使得一个分块内所有单词只属于一个状态,但连续的几个分块内的单词可以属于同一个状态[4]。例如下面一篇论文头部格式为:

Web Mining: Information and Pattern Discovery
on the World Wide Web
R. Cooley, B. Mobasher, and J. Srivastva
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455, USA

如果以换行和逗号作为分块的依据,上面这篇论文头部可以分为 10 个分块(共 31 个单词)。每个分块内的单词只属于一个状态,但连续的几个分块内的单词可以属于同一个状态,如标题状态占两个分块,作者状态占三个分块。在一个分块内部,有些单词属于特征单词,像机构组织中的“Department”,“University”等。而有些单词属于非特征词,像机构组织中的具体名称等。在以单词作为基本抽取单位时,这些非特征词只能依赖转移概率识别,而以分块为抽取单位时,则只要该分块含有一个以上的特征单词,该块就很容易作为一个整体实现正确识别。因此可以提高抽取精确度。

基于文本分块的多模板 HMM 的文本信息抽取算法分为以下四个步骤:

1)文本预处理:对文本进行分块标记。

依据排版格式信息,分隔符等信息将用 HTML 语言标记好的论文头部文本序列转换为由文本分块组成的序列,每一分块都用 HTML 语言进行状态标记。

2)基于马尔可夫链模型,聚类训练数据为几个类[12,13]。

一个马尔可夫过程由其转移概率矩阵和初始概率矩阵定义,其中的转移概率矩阵描述了其动态特性。假设第 k 篇已标记的训练文本分块序列用以下的转移概率矩阵 $\boldsymbol{A}_k$ 表示:

$$\boldsymbol{A}_k = (p_{kij}) = \begin{bmatrix} p_{k11} & \cdots & p_{k1j} & \cdots & p_{k1n} \\ \vdots & & & & \vdots \\ p_{ki1} & \cdots & p_{kij} & \cdots & p_{kin} \\ \vdots & & & & \vdots \\ p_{kn1} & \cdots & p_{knj} & \cdots & p_{knn} \end{bmatrix} \tag{4}$$

其中:

$$p_{kij} = \frac{S_{kij} + \alpha_{kij}}{\sum_{j=1}^{n} (S_{kij} + \alpha_{kij})} \tag{5}$$

其中:

$$\alpha_{kij} = \frac{\beta}{n \times n}$$

S_{kij} 是训练文本分块序列中从标记状态 i 转移到标记状态 j 的次数。通常取 $\beta = n$,n 是模型状态数。

如果第 k 个训练文本分块序列用矩阵 $\boldsymbol{A}_k$ 表示,第 l 个训练文本分块序列用矩阵 $\boldsymbol{A}_l$ 表示,那么这两个训练文本分块序列之间的距离可以用式(6)来计算:

$$d(\boldsymbol{A}_k,\boldsymbol{A}_l)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}p_{kij}\log\frac{p_{kij}}{p_{lij}}+\sum_{i=1}^{n}\sum_{j=1}^{n}p_{lij}\log\frac{p_{lij}}{p_{kij}}}{2\cdot n}\tag{6}$$

在聚类分析中,一般要求定义的距离必须满足非负性和对称性。该距离定义明显满足该要求。对任意两个 Markov 链,当它们具有相同的动态特征时,它们之间的距离为零。它们之间的动态特征差异越大,距离值就越大,这样就能够得到任意两个训练文本分块序列之间的距离。基于这种距离,即可利用层次聚类的方法将训练数据集聚成几个类,通过调节距离的门槛值,可控制得到的聚类个数。

首先,定义类与类之间的距离为其样本之间的最长距离:

$$D_{pq}=\max_{i\in C_p,j\in C_q}d_{ij}\tag{7}$$

然后,基于上述最长距离可得层次聚类的基本算法:

a) 计算样本两两之间距离的对称阵,记为 $\boldsymbol{D}_{(0)}$,开始每个样本自成一类,因此 $D_{pq}=d_{pq}$;

b) 选择 $\boldsymbol{D}_{(0)}$ 的最小元素,设为 D_{pq},将 C_p 与 C_q 合并为一类,记为 C_r,$C_r=\{C_pC_q\}$;

c) 计算新类与其他类之间的距离:

$$D_{rk}=\max_{i\in C_r,j\in C_k}d_{ij}=\max\{\max_{i\in C_p,j\in C_k}d_{ij},\max_{i\in C_q,j\in C_k}d_{ij}\}=\max\{D_{pk},D_{qk}\}$$

将 $\boldsymbol{D}_{(0)}$ 中 p,q 行 p,q 列合并为一个新行新列。新行新列对应 C_r,所得到的矩阵叫作 $\boldsymbol{D}_{(1)}$;

d) 对 $\boldsymbol{D}_{(1)}$ 重复上面对 $\boldsymbol{D}_{(0)}$ 的两步,得到 $\boldsymbol{D}_{(2)}$,如此下去直到所有的元素聚为一类。如果某一步中最小的元素不止一个,则对应这些最小元素的类同时合并。

3) 针对每一类训练数据,都使用 ML 算法来训练一个初始概率矩阵和一个转移概率矩阵,使用式(1) ~ (3)来计算初始概率和转移概率。其中,在式(1)中,$Init(i)$ 是所有以分块为基本单位的训练序列中,初始状态为 i 的序列个数;在式(2)中,$C_{i,j}$是所有以分块为基本单位的训练序列中,从状态 S_i 转换到状态 S_j的次数;在式(3)中,$E_j(V_k)$ 是所有以单词为基本单位的训练序列中(将已标记块序列进一步转化为标记的词序列),状态 S_j 释放单词 V_k 的次数。

因为不同来源的训练文本虽然格式特征上不同,但内容特征还是基本相似的,为了不至于减少训练数据的量,选择将所有的训练文本用来训练一个统一的释放概率矩阵。

4)使用训练好的模型抽取信息时,结合每一个初始概率矩阵、每一个转移概率矩阵和统一的释放概率矩阵,首先使用 Viterbi 算法来找出最优的标记序列。然后,从这些最优标记序列中选择一个概率最大的序列作为最终标记序列。即,对于每一种聚类模板,均首先使用 Viterbi 算法一次,从而使得每一个模板都会产生一个标记序列,然后再从所有这些模板产生的最优标记序列中选出概率最大的序列作为最终输出结果。

3　实验与分析

为了验证算法的性能,下面利用从网上收集的 600 篇论文头部文档进行了仿真实验。具体实验方法如下:以其中 500 篇已标记文本作为训练集,另外 100 篇作为测试集,实验中使用的 HMM 包含 13 个状态。图 1 给出了基于 HMM 的文本信息抽取算法和基于多模板 HMM 的文本信息抽取算法(TBMT-HMM)的对比实验结果。

从图 1 可以看出,使用基于文本分块的多模板隐马尔可夫模型的文本信息抽取算法的精确度比使用单一的隐马尔可夫模型的文本信息抽取算法的精确度高。当训练集增大时,使用基于文本分块的多模板隐马尔可夫模型的信息抽取算法的精确度改善得更为平稳。相比之下,当训练数据集超过 200 篇时,使用单一的隐马尔可夫模型的信息抽取算法的精确度反而下降了。原因可能是增加的训练集和测试集不是很匹配。

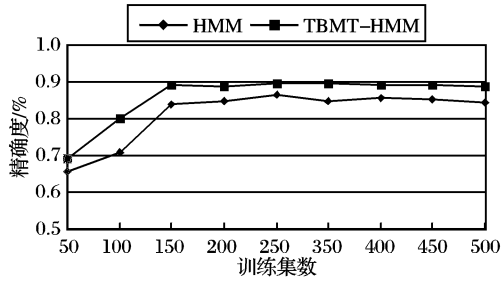

图 1　基于 HMM 与 TBMT-HMM 模型的文本信息抽取精确度比较

由算法描述可知,在基于文本分块的多模板隐马尔可夫模型的文本信息抽取算法中,聚类模板数量的确定非常重要。在实验中,可通过不断改变聚类距离的门槛值,并对训练集本身进行测试,来找到最佳的聚类模板数。图 2 给出了训练集为 400 篇文本时,基于文本分块的多模板隐马尔可夫模型的文本信息抽取算法在不同模板数情况下的文本信息抽取精确度比较。

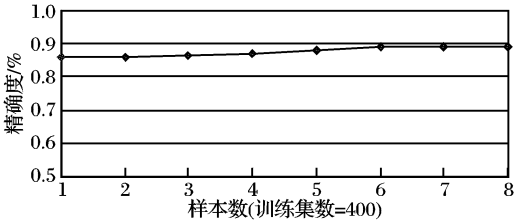

图 2　不同模板数情况下的文本信息抽取精确度比较

由图 2 可知,采用基于文本分块的多模板隐马尔可夫模型的文本信息抽取算法,文本信息抽取的精确度与聚类模板数成正比,即聚类模板数越多,算法的文本信息抽取精确度越高。但当模板数超过 7 时,信息抽取精确度的改进幅度将减弱。再考虑到聚类模板数越多,模型用于信息抽取所耗费的时间也就越多,因此,实验中选择 7 作为最佳的聚类模板数。

表 1　HMM 和 TBMT-HMM 具体的文本信息抽取精确度与召回率比较

Item	HMM		TBMT-HMM	
	精确度/%	召回率/%	精确度/%	召回率/%
Title	0.929 515	0.490 128	0.930 102	0.821 792
Author	0.694 268	0.699 839	0.873 565	0.935 587
Affiliation	0.839 806	0.814 118	0.894 159	0.925 746
Address	0.618 875	0.802 353	0.923 901	0.860 543
Email	0.860 544	0.630 923	0.919 028	1.000 000
Note	0.776 042	0.668 161	0.890 896	0.786 518
Web	1.000 000	0.261 905	1.000 000	0.597 009
Phone	0.942 857	0.407 407	0.988 754	0.943 802
Date	0.636 364	0.954 545	0.738 817	0.992 164
Abstract	0.815 534	1.000 000	0.975 523	1.000 000

有时,由不同数目的聚类所训练得出的模型信息抽取精确度是一样的。一个可能的原因是,对于同一个观察序列来

说,两种不同的训练模型可能会以各自不同的最大概率选择同一条标记序列。另一个可能的原因是,新增加的模板在聚类过程中可能只包括很少的训练文本数,该聚类模板在信息抽取时很少被选中作为最终输出。当训练集为400篇,聚类模板数为7时,各个具体域的抽取精确度与召回率见表1。从中可以看出,基于文本分块的多模板隐马尔可夫模型的抽取精确度和召回率总体上要比基于词的隐马尔可夫模型高,尤其是召回率方面。

4 结语

针对训练数据来源的多样化和难以通过学习得到最优的模型参数,本文提出了一种基于文本分块的多模板隐马尔可夫模型的文本信息抽取算法。新算法首先利用文本排版格式和分隔符等信息,对文本进行分块;然后在此基础上通过对训练数据形式聚类,分多个形式模板训练隐马尔可夫初始概率及转移概率参数;最后,结合统一训练的释放概率参数,对文本信息进行抽取。实验结果表明,新算法在一定情况下能有效提高信息抽取的精确度和召回率。为了进一步提高算法的性能,在后续工作中,我们将定义更合适的模型来描述模板的具体属性,从而使得算法能更加准确地实现训练数据的聚类,以学习到更优的HMM模型参数。

参考文献:

[1] 马亮,陈群秀,蔡莲红. 一种改进的自适应文本信息过滤算法[J]. 计算机研究与发展,2005, 42(1): 79-84.

[2] LIU YI, JIN RONG, JOYCE Y. A maximum coherence model for dictionary-based cross-language information retrieval[C]// Proceedings of the 28 th Annual International ACM SIGIR Conference. Salvador: ACM, 2005: 536-543.

[3] KUSHMERICK N. Wrapper induction: efficiency and expressiveness[J]. Artificial Intelligence Journal, 2000, 118(12): 15-68.

[4] 刘云中,林亚平,陈治平. 基于隐马尔可夫模型的文本信息抽取[J]. 系统仿真学报,2003,16(3): 507-509.

[5] FRIETAG D, McCALLUM A. Information extraction with HMMs and shrinkage[C]// Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction. [S. l.]: IEEE Press, 1999: 31-36.

[6] FREITAG D, McCALLUM A. Information extraction with HMM structures learned by stochastic optimization[C]// Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. AAAI Press/The MIT Press, 2000: 584-589.

[7] RAY S, CRAVEN M. Representing sentence structure in Hidden Markov Models for information extraction[C]// Proceedings of the Seventeenth International Joint Conference On Artificial Intelligence. [S. l.]: IEEE, 2001: 1273-1279.

[8] SCHEFFER T, DECOMAIN C, WROBEL S. Active hidden Markov models for information extraction[C]// Proceedings of the International Symposium on Intelligent Data Analysis. [S. l.]: IEEE Press, 2001: 301-109.

[9] 周顺先. 文本信息抽取模型及算法研究[D]. 长沙:湖南大学, 2007.

[10] RABINER L E. A tutorial on hidden Markov models and selected application in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[11] 林亚平, 刘云中, 陈治平. 基于最大熵的隐马尔可夫模型文本信息抽取[J]. 电子学报,2005,33(2): 236-241.

[12] 邢永康,马少平. 一种基于Markov链模型的动态聚类方法[J]. 计算机研究与发展,2003,40(2): 129-135.

[13] RIDGEWAY G, ALTSCHULER S. Clustering finite discrete Markov chains[C]// Proceedings of the Joint Statistical Meetings, Section on Physical and Engineering Sciences. [S. l.]: IEEE Press, 1998: 228-229.

(上接第687页)

方式,在指定循环次数的条件下设定一阈值,如果符合阈值判断条件则结束循环,并记录循环次数,否则达到最大循环数。表3是对两幅测试图像分别采用原始算法和改进算法进行循环次数和总运算时间的统计结果。

表3 实验数据迭代次数和运算时间统计结果

算法	平均迭代次数	平均运算时间/min
原始算法(测试图像一)	100(设定最大值)	455
改进算法(测试图像一)	32	163
原始算法(测试图像二)	100(设定最大值)	473
改进算法(测试图像二)	27	152

由统计结果可见,改进算法能有效减少在同一温度下系统达到平衡需要循环的次数,从而有效地减少了运算时间。从本文采用的测试图像实验结果来看,速度提高了200%。

4 结语

本文针对传统的MRF-MAP图像分割算法运算量大,运算时间过长的问题提出了一种结合图像的灰度信息的改进算法。首先建立图像灰度分布模型,然后将其用于标号场更新过程中,从而充分的利用了图像的灰度信息,减少了传统模拟退火算法标号场更新时选择不必要的解。实验结果表明改进算法有效地提高了运算速度。

参考文献:

[1] GEMAN S, GEMAN D. Stochastic relaxation, Gibbs distributions, and Bayesian restoration of image[J]. IEEE Transaction PAMI, 1984(6): 721-741.

[2] LI S Z. Markov random field modeling in image analysis springer[M]. Tokyo: Springer-Verlag, 2001.

[3] XIA Y. Adaptive segmentation of textured images by using the coupled Markov random field model[J]. IEEE transactionson image processing[J]. IEEE Transactions on Image Processing, 2006, 15(11): 3559-3566.

[4] DENG H W, DAVID A. Clausi unsupervised image segmentation using a simple MRF model with a new implementation scheme[C]// Proceedings of the 17th international conference on Pattern Recognition. Washington, DC: IEEE Computer Society, 2004: 1051-4651.

[5] MANJUNATH B S. Stochastic and deterministic networks for texture segmentation[J]. IEEE Transactions on acoustics Speech and Singnal Processing, 1990, 38(6): 1039-1049.

[6] RANDASAMI L. Estimation and choice of neighbors in spatial-interaction models of images[J]. IEEE Transactions Information Theory, 1983, IT-29(1): 60-72.

[7] DERIN H, ELLIOTT H. Modeling and segmentation of noisy and textured images using Gibbs random fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1987, 9(1): 39-55.

[8] SENGUR A, TURKOGLU L, INCE M C. Unsupervised image segmentation using Markov random fields[C]// TAINN 205 LNI3949. [S. l.]: ACM Press, 2006: 158-167.