

Markov 逻辑网及其在文本分类中的应用

张玉芳, 黄 涛, 艾东梅, 熊忠阳

(重庆大学 计算机学院, 重庆 400044)

(ram.tnht@gmail.com)

摘 要:介绍了 Markov 逻辑网的理论模型、学习算法和推理算法,并将其应用于中文文本分类中。实验结合了判别式训练的学习算法,MC-SAT、吉布斯抽样和模拟退火等推理算法,结果表明基于 Markov 逻辑网的分类方法能够取得比传统 K 邻近(KNN)分类算法更好的效果。

关键词:统计关系学习;机器学习;Markov 逻辑网;文本分类

中图分类号: TP18; TP391 **文献标志码:** A

Markov logic network and its application in text classification

ZHANG Yu-fang, HUANG Tao, AI Dong-mei, XIONG Zhong-yang

(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: This paper introduced the theory, learning methods and inference algorithms of Markov logic network that was also applied to the Chinese text classification. With reference to the discriminative learning algorithm for Markov logic network weights, MC-SAT, Gibbs sampling and simulated tempering algorithm in experiments, it proves that the method based on Markov logic network is better than conventional K Nearest Neighbor (KNN) method in text classification.

Key words: Statistical Relational Learning (SRL); machine learning; Markov logic network; text classification

统计关系学习(Statistical Relational Learning, SRL)又称概率逻辑学习(Probabilistic Logical Learning, PLL),是人工智能、机器学习和数据挖掘交叉研究的产物,旨在将关系(逻辑)表示、似然推理(不确定性处理)和机器学习(数据挖掘)结合起来,以获取关系数据中的似然模型^[1]。统计关系学习方法由似然关系模型和学习算法组成。似然关系模型是指关系的似然表示形式,学习是指基于数据来调整似然关系模型的过程,即参数学习和结构学习两个任务。统计关系学习的早期研究多集中于归纳逻辑程序设计(Inductive Logic Programming, ILP)^[2-3],随着对 SRL 研究的不断深入,陆续提出许多非 ILP 的统计关系学习方法,Markov 逻辑网就是基于 Markov 网的 SRL 方法。

1 Markov 逻辑网

1.1 Markov 逻辑网简介

Markov 逻辑网^[4]是公式附加权值的一阶逻辑知识库,且可作为构建 Markov 网的模板^[5]。从概率的角度来看,Markov 逻辑网为大型 Markov 网提供一种简洁的描述语言,并能够灵活地将大量领域知识模块化地引入到 Markov 网中;从一阶逻辑的角度来看,Markov 逻辑网不仅可以处理不确定性,还可以允许不完整和矛盾的知识。此外,Markov 逻辑网还可以作为很多统计关系学习任务的统一框架^[6]。

在一阶逻辑中,通常一个世界只要违反了一个公式,该世界发生的概率就为 0(即一阶逻辑知识库作为可能世界的强约束)。Markov 逻辑网的基本思想是软化这个约束:当一个世界违反了知识库中的一个公式,该世界发生的概率降低但不是不可能;而违反的公式越少,则发生的概率越大。公式上

的权值体现了该公式的限制强度,权值越大,满足该公式的世界的发生概率与不满足该公式的世界的发生概率之间的差就越大。随着公式上权值的增加,Markov 逻辑网逐渐向纯逻辑知识库靠拢。

定义 1 Markov 逻辑网 $L^{[4]}$ 为二元组集合 $\{(F_i, w_i)\}$, 其中 F_i 为一阶逻辑表示的公式, w_i 为实数。给定 Markov 逻辑网 L 和有限个体常项集合 $C = \{C_1, C_2, \dots, C_{|C|}\}$, 则可生成一个以闭谓词为节点、闭谓词关系为边的 Markov 网 $M_{L,C}$:

1) L 中每个闭谓词对应 $M_{L,C}$ 中的一个节点,若闭谓词为真,则节点值为 1,否则为 0;

2) L 中每个公式 F_i 的闭公式对应 $M_{L,C}$ 中一个特征函数,若闭公式为真,则特征函数值为 1,否则为 0。特征函数值就是 L 中 F_i 所对应的 w_i 。为了计算方便,通常将势函数 ϕ_k 定义为:

$$\phi_i(x_{|i|}) = \begin{cases} e^{w_i}, & F_i \text{ 为真} \\ 1, & \text{否则} \end{cases} \quad (1)$$

对于一个 Markov 逻辑网,给定不同的个体常项集合,就产生不同的 Markov 网。由于给定的个体常项集合规模不同(个体常项多,则生成 Markov 网的节点就多,网的规模也就越大),生成的 Markov 网的规模可能完全不同。但是因为它们来自于同一个 Markov 逻辑网,所以其参数相同且结构有相同的规律性(参数相同,是因为对应同一公式的所有闭公式均拥有相同权值),称这些 Markov 网为闭 Markov 逻辑网。根据 Markov 网的定义,闭 Markov 逻辑网 $M_{L,C}$ 的概率分布应为:

$$P(X = x) = \frac{1}{Z} \exp\left[\sum_i w_i n_i(x)\right] = \frac{1}{Z} \prod_i \phi_i(x_{|i|})^{n_i(x)} \quad (2)$$

其中: X 为一个世界中所有可能闭谓词所构成的向量, x 表示

收稿日期:2009-04-02;修回日期:2009-05-17。 基金项目:重庆市自然科学基金资助项目(CSTC 2008BB2021)。

作者简介:张玉芳(1965-),女,上海人,副教授,主要研究方向:数据挖掘、网络入侵检测; 黄涛(1982-),男,安徽合肥人,硕士研究生,主要研究方向:统计关系学习、机器学习、数据挖掘; 艾东梅(1985-),女,四川内江人,硕士研究生,主要研究方向:数据挖掘、语义网; 熊忠阳(1962-),男,重庆人,教授,博士生导师,主要研究方向:网络与并行处理、数据挖掘、互联网。

闭谓词的真值, $Z = \sum_{x \in X} \prod_k \phi_k(x_{|k|})$ 为分配函数、以保证 $\sum_{x \in X} P(X = x) = 1$, w_i 即式(1)中给出的权值, $n_i(x)$ 表示与世界 x 相对应的公式 F_i 的真闭公式个数, $x_{|i|}$ 表示出现在公式 F_i 中闭谓词的真值, $\phi_i(x_{|i|}) = e^{w_i}$ 。理论上, $n_i(x)$ 和 $\phi_i(x_{|i|})$ 都可以通过给定的知识库计算得到。

1.2 Markov 逻辑网的参数学习算法

对于 Markov 逻辑网而言, 参数是知识库中各公式的权值 $w_i (i = 1, \dots, m)$, 因此参数学习任务就是估计出知识库中所有公式的权值。先假设世界是完备的^[14]: 若一个闭谓词不在数据库中, 则该闭谓词为假。在完备世界假设下, 若有 n 个可能闭谓词, 一个数据库就是一个向量 $\mathbf{x} = (x_1, \dots, x_l, \dots, x_n)$, 其中 x_l 为第 l 个闭谓词的真值, 若该闭谓词在数据库中出现, 则 $x_l = 1$; 否则, $x_l = 0$ 。

给定一个数据库, Markov 逻辑网的权值原则上可以通过最大似然估计的方法学习到。即参数 w_i 看作固定值, 并假设所有数据满足参数 w_i , 通过计算使 $X = \mathbf{x}$ 的似然概率 $P_w(X = \mathbf{x})$ 取最大值的 $w_i (i = 1, \dots, m)$ 来获取参数值。如果与世界 \mathbf{x} 相对应的第 i 个公式有 $n_i(\mathbf{x})$ 个真闭公式, 则根据式(2)可以导出式(3):

$$\frac{\partial}{\partial w_i} \log P_w(X = \mathbf{x}) = n_i(\mathbf{x}) - \sum_{\mathbf{x}'} P_w(X = \mathbf{x}') n_i(\mathbf{x}') \quad (3)$$

其中, $n_i(\mathbf{x})$ 和 $n_i(\mathbf{x}')$ 分别表示与世界 \mathbf{x} 和任意世界 \mathbf{x}' 相对应的公式 F_i 的真闭公式个数。令式(3)等于 0 并求解 $w_i (i = 1, \dots, m)$ 需要知道 $n_i(\mathbf{x})$ 和 $n_i(\mathbf{x}')$ 的值, 理论上 $n_i(\mathbf{x})$ 和 $n_i(\mathbf{x}')$ 都可以从数据库中计算得到。由于在单一数据库中计算公式的真闭公式个数是 NP 问题^[4], 故对于实际数据库无法通过计算 $n_i(\mathbf{x})$ 和 $n_i(\mathbf{x}')$ 来求解。最大似然估计不可行, 则可以考虑用最大伪似然估计方法 (Maximum Pseudo-Likelihood Estimation, MPLE) 和判别式训练方法 (Discriminative Training) 来替代^[7]。

最大伪似然估计方法用伪似然概率替代似然概率, 即:

$$\frac{\partial}{\partial w_i} \log P_w^*(X = \mathbf{x}) = \sum_{l=1}^n [n_i(\mathbf{x}) - P_w(X_l = 0 | MB_x(X_l)) \times n_i(\mathbf{x}_{[X_l=0]})] - P_w(X_l = 1 | MB_x(X_l)) \times n_i(\mathbf{x}_{[X_l=1]}) \quad (4)$$

其中: $P_w^*(X = \mathbf{x})$ 为伪似然概率, $MB_x(X_l)$ 表示 X_l 的 Markov 覆盖^[8], $n_i(\mathbf{x}_{[X_l=0]})$ 表示指定 $X_l = 0$ 时第 i 个公式的真闭公式个数, $n_i(\mathbf{x}_{[X_l=1]})$ 含义类似。由于 $P_w(X_l = 0 | MB_x(X_l))$ 和 $P_w(X_l = 1 | MB_x(X_l))$ 可从数据库中直接得到, 故 $P_w(X_l = 0 | MB_x(X_l))$, $P_w(X_l = 1 | MB_x(X_l))$, $n_i(\mathbf{x}_{[X_l=0]})$ 和 $n_i(\mathbf{x}_{[X_l=1]})$ 均可通过计算得到, 因此当上式等于 0 时, 参数学习问题转化为非线性优化问题。但是最大伪似然估计方法会导致非邻接变量之间的推理结果不理想。为了解决该问题, 可以采用判别式训练方法。

在许多应用中, 如果我们事先知道哪些谓词是证据谓词哪些谓词是查询谓词, 则在给定证据谓词的条件下可以通过推理来正确预测查询谓词。判别式训练方法就是把域中的闭原子分为两个集合: 证据原子集合 X 和查询原子集合 Y , 在给定 X 的条件下, Y 的条件是:

$$P(y | x) = \frac{1}{Z_x} \exp \left(\sum_{i \in F_Y} w_i n_i(x, y) \right) = \frac{1}{Z_x} \exp \left(\sum_{j \in G_Y} w_j g_j(x, y) \right) \quad (5)$$

其中: Z_x 为给定 X 的条件下的分配函数, F_Y 是所有 Markov 逻辑网的子句 (至少有一个闭子句涉及到查询原子) 所构成的集合; $n_i(x, y)$ 是涉及到查询原子的第 i 个子句的真闭子句的个数; G_Y 是 $M_{L,C}$ 中涉及到查询原子的闭子句集合; 当数据中第 j 个闭子句为真时, $g_j(x, y)$ 的值为 1, 否则 $g_j(x, y)$ 的值为 0。当隐藏某些变量时 (既非查询谓词, 也非证据谓词), 可通过累加求和来计算条件似然。为了叙述简便, 将所有的非证据变量都看作查询变量。式(5)微分后可通过计算 $n_i(x, y_w^*)$ 来求其近似值 ($y_w^*(x)$ 表示 MAP (Maximum A Posteriori) 状态), 于是式(5)的计算就可以转化为用 MAP 推理来寻找 $y_w^*(x)$ 。

此外, 基于最大伪后验估计的 Markov 逻辑网参数学习方法^[12-13]同样可以取得不错的学习效果。

1.3 Markov 逻辑网的概率推理算法

Markov 逻辑网可以回答任意类似“给定公式 F_1 的情况下公式 F_2 成立的概率是多少?”这样的查询问题。如果 F_1 和 F_2 是两个一阶逻辑公式, C 表示出现在 F_1 和 F_2 中个体常项的有限集合, L 表示 Markov 逻辑网, 则有:

$$P(F_1 | F_2, L, C) = P(F_1 | F_2, M_{L,C}) = \frac{P(F_1 \wedge F_2 | M_{L,C})}{P(F_2 | M_{L,C})} = \frac{\sum_{x \in \mathcal{X}_{F_1} \cap \mathcal{X}_{F_2}} P(X = x | M_{L,C})}{\sum_{x \in \mathcal{X}_{F_2}} P(X = x | M_{L,C})} \quad (6)$$

其中, \mathcal{X}_{F_i} 表示使公式 F_i 成立世界的集合, $P(x | M_{L,C})$ 由式(2)给出。

在图模型中, 常见的条件查询是式(6)的特殊情况, 即所有 F_1, F_2 和 L 中的谓词都是零元的且公式都是合取的。此时, 问题“知识库 KB 中是否包含公式 F ”等价于判断 $P(F | L_{KB}, C_{KB,F}) = 1$ 是否成立的问题。其中, L_{KB} 是 KB 中所有公式附加权值所得到的 Markov 逻辑网, $C_{KB,F}$ 是出现在 KB 或 F 中的所有个体常项。上述问题的答案即在 $F_2 = \text{True}$ 的条件下, 通过式(6)来计算 $P(F | L_{KB}, C_{KB,F})$ 。

由于 Markov 逻辑网的推理是 NP-完全问题^[5], 即便是在有限的领域进行推理, 也很难有好的结果。原则上, 计算 $P(F_1 | F_2, L, C)$ 可以近似地使用马尔可夫链蒙特卡洛方法 (Markov Chain Monte Carlo, MCMC), 在 F_2 不成立时不做任何改变, 在 F_1 成立时计算抽样数。

MCMC^[11] 方法是一种特殊的蒙特卡洛方法, 它将随机过程中的马尔可夫过程引入到蒙特卡洛模拟中, 实现动态模拟 (即抽样分布随模拟的过程而改变)。本质上, MCMC 方法是使用马尔可夫链的蒙特卡洛积分。蒙特卡洛积分是通过抽样点 $\{X^{(t)}, t = 0, 1, \dots\}$ 来估计函数 $h(X)$ 的期望, 其估算公式为: $E[h(X)] = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n h(X^{(i)})$ 。这样, 通过估计 $h(X)$ 的均值可以得到总体的均值, 当抽样点 $X^{(t)}$ 相互独立时, 可以增加抽样次数 n 来提高估计精度, 并且经过一段时间的迭代, $X^{(t)}$ 的分布可以收敛到一个平稳分布。这时, MCMC 算法的估算式应当去掉收敛以前的迭代, 用收敛后的迭代值来估计。由于 MCMC 方法的基本思想是通过建立一个平稳分布为 $\pi(x)$ 的 Markov Chain 来得到 $\pi(x)$ 的样本。因此, 构造转移核以使已知的概率分布 $\pi(x)$ 为平稳分布是至关重要的。不同的转移核将导致不同的 MCMC 方法, 本文所用到的有 MC-SAT、吉布斯抽样、模拟退火三种算法。吉布斯抽样和模拟退火是

MCMC 方法中常用的简单变体,这里不再赘述,下面简要介绍下 MC-SAT 算法。

MC-SAT 是一种切片抽样的 MCMC 算法,切片抽样是一种通过一个预先定义好的不变分布来构造可逆马尔可夫转移核的方法,实际上是一种辅助变量法。MC-SAT 联合使用了可满足性测试和模拟退火,使用可满足性求解器(WalkSAT)的好处是可以有效地找到分布中的分离模式,因而在马尔可夫链中收敛很快。较之标准 MCMC 方法,如吉布斯抽样和模拟退火,MC-SAT 算法的收敛速度要快很多。这里给出 MC-SAT 算法的伪码表示:

```
MC-SAT( clauses, weights, num_samples )
 $x^{(0)} \leftarrow \text{Satisfy}(\text{hard clauses})$ 
for  $i \leftarrow$  to num_samples do
   $M \leftarrow \emptyset$ 
  for all  $c_k \in \text{clauses}$  satisfied by  $x^{(i-1)}$  do
    With probability  $1 - e^{-w_k}$  add  $c_k$  to  $M$ 
  end for
  Sample  $x^{(i)} \sim U_{\text{SAT}(M)}$ 
end for
```

2 文本分类

传统的文本分类步骤如下:1) 预处理;2) 将文档表示成算法易于处理的向量形式;3) 特征词选取和权值计算;4) 用已经训练好的分类器给未知类别的文档分类;5) 评估分类效果。预处理阶段主要是去除停用词、低频词以及网页文本中的各类标记信息。若处理的是中文文本,则在预处理的时候还要事先进行分词。常用的文本表示形式有向量空间模型和正交分解模型^[10]。通常,选择特征词是本着用最少的词最贴切地反映文档主题的原则,常见的方法有:文档频率方法、信息增益方法、互信息方法以及 χ^2 统计量等。经过特征选取后,将特征的维数降低到一个合理的范围内,就可利用分类器进行分类处理了。使用较广泛的分类方法有:朴素贝叶斯、支持向量机、K 最近邻分类算法(K-Nearest Neighbor, KNN)、决策树等。

由于文本分类的任务本质上是判断某个类是否有某篇文本,而通常采取的特征是文本的基本单位——“词”,也就是某篇文本有某些词。据此,可以将文本分类用两个谓词表示,即:HasWord(word, text) 表示文本 text 中有词 word; Topic(class, text) 表示文本 text 属于 class 类。

根据这两个谓词,很容易可以得到公式 HasWord(word, text) = > Topic(class, text), 表示某文档 text 中有词 word, 则该文档属于类 class。由于针对同一文档而言,有不同的词,且属于不同类别的概率也不尽相同。在 MLN 表示中,当公式中的变量之前加上“+”号,每个公式依据变量所代表的个体不同,分别学习到不同的权值。所以可在 .mln 文件中将该公式写为 HasWord(+w, p) = > Topic(+c, p)。表示针对同一篇文档 t 而言,分别对不同的词 w 和不同的类 c 学习不同的权值。

3 实验及分析

3.1 数据集

本文实验的目的是检验 Markov 逻辑网在中文文本分类中的效果,实验采用复旦大学李荣陆博士的中文分类语料的一个小型语料库。共 2816 篇文档,从中随机抽取了 1882 篇文档作为训练集,934 篇作为测试集,共分电脑、艺术、教育、交通、环境、经济、医疗、军事、政治、体育 10 个类。

3.2 实验方法

实验预处理在 Microsoft Visual Studio 2005 下进行, Markov 逻辑网相关的权值学习和概率推理在 Alchemy^[9] 下进行。Alchemy 是 Domingos 等人开发的基于 Markov 逻辑表示的软件包,提供了一系列统计关系学习和概率逻辑推理方面的算法。实验具体过程分为以下几个步骤。

1) 文本分类的 Markov 逻辑表示,即构建 .mln 文件,内容如下:

```
HasWord( word, text )
Topic( class, text )
HasWord( +w, p ) = > Topic( +c, p )
```

2) 中文文本分词。文本分词采用中国科学院 ICTCLAS 2009 分词程序的 C# 版本。

3) 建立索引。索引采用开源 Lucene 程序的 C# 版本,针对分词后的文本建立倒排索引。训练集共有词 39 900 个,测试集共有词 25 600 个。

4) 特征提取。特征提取采用 χ^2 统计量,并选取 5%、15%、25%、50%、75%、100% 等不同数量的特征词作为构建知识库所需的谓词输出。

5) 知识库构建。学习所用的知识库需要提供 HasWord(word, text) 和 Topic(class, text) 两种类型的闭谓词,根据不同数量的特征词,分别生成不同的 .db 文件。

推理所用的知识库仅需要 HasWord(word, text) 类型的闭谓词,然后通过推理得到 Topic(class, text) 谓词,从而判断某文本是否的确属于某类别。

6) 权值学习。本文采用的权值学习算法为判别式训练方法,针对不同数量的特征词分别学习其权值。

7) 谓词推理。根据上一步学习到的权值和测试知识库,我们采用了三种 MCMC 算法进行推理。

3.3 实验结果及分析

对文本分类的单个类的分类效果常采用查全率和查准率来评价。其数学公式表示如下:

$$\text{查全率}(R) = \frac{\text{分类正确的文本数}}{\text{应有的文本数}}$$

$$\text{查准率}(P) = \frac{\text{分类正确的文本数}}{\text{实际分类的文本数}}$$

$$F1 \text{ 值}(F1) = \frac{P \times R \times 2}{P + R}$$

为了从全局的角度评价分类性能,有两种方法来综合所有类别的分类情况,即宏平均和微平均。宏平均更多地受到稀有类的影响,而微平均更看重重大类别的分类结果,下面给出两种评价方法对应的查全率、查准率、F1 值的数学公式(其中 C 为类别集合, c 为某一类别, a 为某类别中分类正确的文本数, b 为某类别中分类错误的文本数):

$$\text{Macro-R} = \frac{\sum_{i=1}^{|C|} R_i}{|C|}, \text{Macro-P} = \frac{\sum_{i=1}^{|C|} P_i}{|C|},$$

$$\text{Macro-F1} = \frac{2 \times \text{Macro-R} \times \text{Macro-P}}{\text{Macro-R} + \text{Macro-P}},$$

$$\text{Micro-R} = \frac{\sum_{i=1}^{|C|} a_i}{\sum_{i=1}^{|C|} a_i + \sum_{i=1}^{|C|} b_i}, \text{Micro-P} = \frac{\sum_{i=1}^{|C|} a_i}{\sum_{i=1}^{|C|} a_i + \sum_{i=1}^{|C|} c_i},$$

$$\text{Micro-F1} = \frac{2 \times \text{Micro-R} \times \text{Micro-P}}{\text{Micro-R} + \text{Micro-P}}$$

表 1 给出了不同推理算法下,分别针对不同数量特征词的宏平均查全率、宏平均查准率、宏平均 $F1$ 值、微平均查全率、微平均查准率和微平均 $F1$ 值。图 1 给出了与 KNN 查准率的比较结果。

表 1 不同算法、不同特征数下分类效果比较

推理算法	性能指标	特征数比例/%					
		5	15	25	50	75	100
MC-SAT	Macro-R	0.861 65	0.879 84	0.888 34	0.888 57	0.893 62	0.893 46
	Macro-P	0.900 79	0.916 68	0.922 12	0.920 53	0.928 87	0.924 36
	Macro-F1	0.880 79	0.897 88	0.904 92	0.904 27	0.910 90	0.908 64
	Micro-R	0.875 67	0.894 96	0.901 39	0.902 47	0.906 85	0.905 78
	Micro-P	0.875 67	0.894 96	0.901 39	0.902 47	0.906 85	0.905 78
	Micro-F1	0.412 23	0.432 87	0.435 26	0.434 78	0.434 78	0.434 06
吉布斯抽样	Macro-R	0.863 33	0.879 61	0.886 20	0.896 51	0.898 34	0.897 32
	Macro-P	0.904 77	0.917 21	0.922 30	0.931 12	0.930 71	0.928 32
	Macro-F1	0.883 56	0.898 01	0.903 89	0.913 48	0.914 24	0.912 56
	Micro-R	0.877 81	0.894 96	0.899 25	0.908 90	0.910 06	0.908 99
	Micro-P	0.877 81	0.894 96	0.899 25	0.908 90	0.910 06	0.908 99
	Micro-F1	0.413 79	0.431 69	0.431 69	0.432 43	0.433 24	0.432 88
模拟退火	Macro-R	0.863 29	0.883 71	0.886 46	0.899 26	0.896 37	0.898 54
	Macro-P	0.903 99	0.919 97	0.922 36	0.932 03	0.929 71	0.929 03
	Macro-F1	0.883 17	0.901 48	0.904 06	0.915 35	0.912 74	0.913 53
	Micro-R	0.877 81	0.898 18	0.899 25	0.911 40	0.908 99	0.910 06
	Micro-P	0.877 81	0.898 18	0.899 25	0.911 40	0.908 99	0.910 06
	Micro-F1	0.416 00	0.432 88	0.430 52	0.433 60	0.435 97	0.434 07

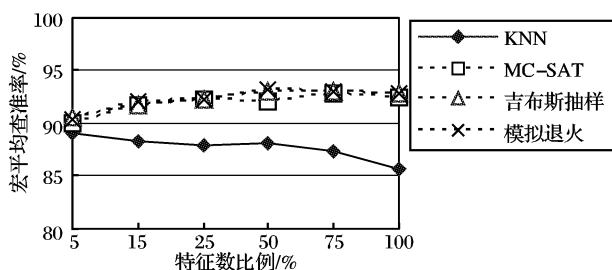


图 1 Markov 逻辑网与 KNN 查准率对比图

基于以上数据,我们得出以下结论:

1) 采用吉布斯抽样和模拟退火算法的分类效果要优于 MC-SAT,这是因为 MC-SAT 算法在提高收敛速度的同时损失了一部分精度;

2) 本文采用的基于 Markov 逻辑网的文本分类,效果明显好于传统的 KNN 分类算法;

3) 推理时,当特征数量增加到 100% 时,其效果稍有下降,但不明显,下降的原因是由于构建的 Markov 逻辑网过大而导致;

4) 当特征数量增加到 100% 时,KNN 分类效果明显下降,而 Markov 逻辑网却仍然有很好的分类效果。这是因为在权值学习时,Markov 逻辑网会为每个词所在的公式赋予权值,若该词对分类贡献较少,则学到的权值也小,这也表明了 Markov 逻辑网在不提取特征的情况下也能获取很好的分类效果,对于干扰词有较大的耐受程度。

4 结语

本文将 Markov 逻辑网应用到中文文本分类中,并通过实验证实了该方法的有效性。在统计关系学习中,可以通过逻辑(关系)来很好地表示知识,故文本分类问题的 Markov 逻辑表示也十分简洁。针对其他不同的应用领域,若存在某些可以用逻辑表示的关系,同样可以采用 Markov 逻辑网来解决问题。比如在超文本分类中,传统做法都是将其当作文本来处理,忽略了页面之间的链接关系,如何利用页面间的链接关系

来提高分类效果,这也是本文的下一步工作。

参考文献:

- [1] RAEDT L D, KERSTING K. Probabilistic logic learning[J]. ACM-SIGKDD Explorations: Special Issue on Multi-Relational Data Mining, 2003, 5(1): 31-48.
- [2] DZEROSKI S. Relational data mining[M]. [S. l.]: Springer, 2005: 869-898.
- [3] DZEROSKI S. Multi-relational data mining: an introduction[J]. ACM SIGKDD Explorations Newsletter, 2003, 5(1): 1-16.
- [4] RICHARDSON M, DOMINGOS P. Markov logic networks[D]. Seattle, Washington, USA: University of Washington, 2004.
- [5] RICHARDSON M, DOMINGOS P. Markov logic networks[J]. Machine Learning, 2006, 62: 107-136.
- [6] DOMINGOS P, RICHARDSON M. Markov logic: A unifying framework for statistical relational learning[C]// Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields. New York: ACM Press, 2004: 49-54.
- [7] SINGLA P, DOMINGOS P. Discriminative training of Markov logic networks[C]// Proceedings of 20th National Conference on Artificial Intelligence. PA: AAAI Press, 2005: 868-873.
- [8] PEARL J. Probabilistic reasoning in intelligent systems: Networks of plausible inference[M]. San Francisco: Morgan Kaufmann Publishers, 1988.
- [9] KOK S, SINGLA P, RICHARDSON M, et al. The alchemy system for statistical relational AI[R]. Seattle, WA: University of Washington, Department of Computer Science and Engineering, 2005.
- [10] 熊忠阳, 李智星, 张玉芳, 等. 一种基于正交分解的文本分类新模型[J]. 计算机工程, 2008, 35(14): 227-229.
- [11] ANDRIEU C, de FREITAS N, DOUCET A, et al. An introduction to MCMC for machine learning[J]. Machine Learning, 2003, 50: 5-43.
- [12] 孙舒杨. 统计关系学习的若干问题研究[D]. 长春: 吉林大学, 2006.
- [13] 孙舒杨, 刘大有, 孙成敏. 基于后验概率的 Markov 逻辑网参数学习方法[J]. 吉林大学学报: 理学版, 2006, 44(6): 216-217.
- [14] GENESERETH M R, NILSSON N J. Logical foundations of artificial intelligence[M]. Los Altos, CA: Morgan Kaufman Publishers, 1987: 406.