基于低秩约束的稀疏主题模型

刘 超 庄连生 俞能海
(中国科学技术大学信息学院 合肥 230027)

摘 要 传统潜在语义分析模型所得到的主题空间映射矩阵往往比较稠密,不仅存储代价比较高,而且各个主题含义不明确。针对该问题,提出一种新的稀疏主题模型,该模型通过对映射矩阵施加稀疏性约束,使得每个主题只与少数词项关联,来增加主题的可解释性;同时,通过对编码系数矩阵施加低秩约束,使得数据在主题空间中呈现出更好的聚类特性。实验结果表明,基于该模型得到的主题空间更有利于分类,映射矩阵的存储代价更低。

关键词 主题模型,稀疏表示,低秩表示
中图法分类号 TP391.4A **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.10.021

STMLRC:Sparse Topic Model with Low Rank Constraint

LIU Chao ZHUANG Lian-sheng YU Neng-hai
(School of Information, University of Science and Technology of China, Hefei 230027, China)

Abstract The project matrix learned by classic Latent Semantic Analysis is always dense, which leads to high storage cost and unclear semantic for each topic. To tackle this problem, a novel sparse topic model was proposed in this paper. By enforcing the sparsity of project matrix, the new model only selects a small number of relevant words for each topic and hence leads to a clear semantic interpretation. Moreover, by enforcing the low rankness of encoding matrix, data projected in the topic subspace shows a better clustering features. Experimental result show that topic subspace learned by our new topic model is in favor of classification, and significantly reduces the storage cost of project matrix.

Keywords Topic model, Sparse representation, Low rank representation

1 引言

潜在语义分析(Latent Semantic Analysis, LSA)[1,2]是目前使用最为广泛的一个知识表示和数据降维方法,在文本挖掘、自然语言处理、信息检索等领域中有着广泛应用。不同于传统的空间向量模型,LSA假设文本中词与词之间存在着某种联系,即某种潜在的语义结构。这种潜在的语义结构不以词典上的定义为基础,而是隐含在词语的上下文使用环境中,可以通过对大量文本集进行统计分析而获取。每种潜在语义结构对应一个主题,所有的潜在语义结构组成了文档的主题空间。文档是主题空间中的一个表示(而在传统空间向量模型中,文档是词典空间中的一个表示)。LSA的目的是根据语料集(即文本集)寻找一个主题空间投影矩阵,通过将文档从词典空间映射到主题空间,去除原始文档中的数据噪声,明确文档的语义关系,消除词语相关性,实现文档的低维表示。

奇异值分解(SVD)是LSA构造主题空间投影矩阵普遍采用的一种方法。设 $D\in R^{N\times M}$ 是给定的一个"文档-词项"共生矩阵,N 为文本集中文档个数,M 为词典中词项个数,LSA首先对 D 进行矩奇异值分解(SVD),得到 $D\approx USV^T$。其中,矩阵 $U\in R^{N\times T}$ 和矩阵 $V\in R^{M\times T}$ 分别是由 D 左特征向量和右特征向量构成的矩阵,S 是由奇异值构成的对角阵。然后,LSA取前 K 个最大奇异值及其对应的特征向量构成新的对角阵 S_K 和右特征向量矩阵 V_K。这里 K 是降维之后主题空间的维数。最后,投影矩阵被定义为 $A=S^{-1}V^T$。该方法虽然计算简单,但是却存在着如下几个重要缺陷:(1)利用SVD分解构建的投影矩阵非常稠密,使得每个主题与所有词语相关联,导致主题不够明确。事实上,虽然词典规模比较大,但是每个词语所能表达的语义含义相对有限,因此每个主题只与少数词语相关才是比较合理的,也就是投影矩阵应该稀疏;(2)SVD只能处理高斯噪声,对语料集质量要求非常高。在实际应用中,文档中经常包含一些和所描述语义毫不相关的词项,形成稀疏损毁噪声,极大降低LSA算法性能;(3)稠密的投影矩阵也增加了算法的存储代价和计算复杂度。

为了解决上述问题,Xi Chen等人提出一种稀疏潜在语义分析模型(Sparse Latent Semantic Analysis, SLSA)[3],它把主题空间学习问题转化为矩阵分解问题,通过最小化矩阵的L1范数来对投影矩阵施加稀疏性约束。与传统的LSA模型相比,SLSA可以自动选择最相关的词语来表达每个主题(而不是利用所有词语),增加了主题的可解释性。随着投影矩阵变得稀疏,矩阵的存储代价和投影变换的计算复杂度也

到稿日期:2013-07-05 返修日期:2013-08-19

刘 超(1987-),男,硕士生,主要研究方向为计算机视觉、信息检索,E-mail:lccu@126.com;庄连生(1978-),男,博士,讲师,主要研究方向为计算机视觉、图像检索、机器学习,E-mail:lszhuang@usts.edu.cn(通信作者);俞能海(1964-),男,博士,教授,博士生导师,主要研究方向为多媒体信息检索。

随之降低。

从根本上说，主题模型(包括 LSA)的目标在于为文档数据寻找一个新的低维主题空间，以方便进行各种数据分析(如数据聚类)。从文档分类的角度来看，我们不仅希望主题空间具有更明确的语义含义(即可解释性)，还希望文档数据映射到主题空间后更有利于分析，呈现出更好的聚类特性。然而，现有 SLSA 算法虽然较好地解决了投影矩阵的稀疏性问题，但是却忽略了样本在主题空间中的分布问题。这在一定程度上降低了算法的性能，使得 SLSA 还有较大的提升空间。

基于上述分析，本文提出一种新的主题模型——基于低秩约束的稀疏主题模型(Sparse Topic Model with Low Rank Constraint，STMLRC)，简称为低秩稀疏主题模型。和 SLSA 类似，STMLRC 模型把主题空间学习问题转化为矩阵分解问题，借助于稀疏表示技术对主题空间映射矩阵施加稀疏性约束，保证每个主题只与少数词项相关联，使得各个主题的语义含义更加明确。同时，为了保证主题空间更加有利于分类，STMLRC 方法对训练样本在主题空间中的编码系数施加低秩约束，以保证映射后同类样本可以更好地聚类在一起。利用低秩约束来实现全局约束已经在文献[4-6]中得到证明，本文工作主要受到这些工作的启发。最后，本文借助于增广拉格朗日算法(ALM)[7]来对 STMLRC 模型进行求解。在公共数据集上的实验结果表明，相比于最新的 SLSA 模型，STMLRC 模型得到的主题空间具有更好的分类性能。

本文第 2 节给出 STMLRC 模型及其相关求解算法；第 3 节是实验结果和分析；最后对全文做总结。

2 STMLRC 模型及其求解算法

2.1 低秩稀疏主题模型

和传统 LSA 不同，STMLRC 方法把主题空间学习问题转化为矩阵分解问题。对于给定的文档-词项共生矩阵 $D\in R^{M\times N}$，其中 N 为文档总数，M 为词项总数。假设隐含的主题空间的维数为 $T(T\leqslant\min\{N,M\})$，则 STMLRC 模型把 D 表示为 $A\in R^{M\times T}$ 和 $X\in R^{T\times N}$ 之积，即 $D=AX$。其中，A 就是主题空间映射矩阵，X 为观测文档(即文本集)在主题空间中的映射。A 中每一列对应着一个主题，用词典向量来表示。为了保证可解释性，STMLRC 对 A 施加稀疏性约束，以保证主题每个主题只与少数词项相关。同时，为了使得主题空间具有更好的分类性能，STMLRC 对 X 施加低秩约束，要求 X 呈现出加强的聚类特性。为此，STMLRC 模型求解如下问题：

$$\begin{cases}\min \lambda\cdot\|A\|_0+\beta\cdot rank(X)\\ \text{s.t. } D=AX\end{cases}\tag{1}$$

式中，$\|\cdot\|_0$ 表示矩阵 A 的 0 范数，定义为 A 矩阵中非零元素的和；$rank(X)$ 代表矩阵 X 的秩；λ 和 β 是对应的系数，分别控制了稀疏性和低秩约束在模型中的重要性。对字典矩阵 A 进行 0 范数约束是为了使其具有稀疏性，如果 A 中的非零元素个数不多于 k 个($k<<T$)，就称矩阵 A 是 k 稀疏的；在实际应用中，如果 A 至多只有 k 个比较明显的非零元素，而其他的非零元素相对较小(接近于零)，那么也称 A 是 k 稀疏的。对编码矩阵 X 的秩求最小值，是为了实现对 X 的低秩约束。显然，问题(1)是一个 NP 难的非凸优化问题，难以求解。幸运的是，根据压缩感知和矩阵恢复相关理论[8,9]，如果 A 足够稀疏，我们通常可以用 l_1 范数代替 0 范数，而矩阵秩的最小化问题可以转化为求其最小化的核范数。因此，STMLRC 模型把问题(1)转化为如下凸优化问题：

$$\begin{cases}\min \lambda\|A\|_1+\beta\|X\|_*\\ \text{s.t. } D=AX\end{cases}\tag{2}$$

式中，$\|\cdot\|_1$ 表示 l_1 范数，其定义为矩阵所有元素的绝对值的和；$\|\cdot\|_*$ 表示核范数，其定义为其矩阵奇异值的和。在实际的应用中，文档集中经常包含一些和主题不相关的词项。这可以通过稀疏噪声来进行建模。为此，STMLRC 模型对问题(2)进行修正：

$$\begin{cases}\min \lambda\|A\|_1+\beta\|X\|_*+\|E\|_1\\ \text{s.t. } D=AX+E\end{cases}\tag{3}$$

显然，问题(3)需要同时求解出最优的 A 和 X，并非是一个凸优化的问题。但是当固定 A、X、E 中任意两个变量时，目标方程对于剩余那个变量则为可求解的凸优化问题。为此，本文拟采用增广拉格朗日算法(ALM)来求解。

2.2 基于 ALM 的求解算法

增广拉格朗日算法是目前常用的一种高维凸优化问题求解算法，算法求解简单，速度快。问题(3)的增广拉格朗日函数为：

$$L(A,X,E,Y,u)=\lambda\|A\|_1+\beta\|X\|_*+\|E\|_1+\langle Y,D-AX-E\rangle+\frac{\mu}{2}\|D-AX-E\|_F^2\tag{4}$$

式中，$\langle Y,D-AX-E\rangle$ 表示矩阵点乘。根据 ALM 算法，各个变量按照如下规则来更新：

1. 固定 X,E,Y,u，更新 A：

$$\begin{aligned}A_{k+1}&=\arg\min_A \lambda\|A_k\|_1+\frac{\eta_A u_k}{2}\|A-A_k+[-(D-A_kX-E_k+\frac{Y_k}{u})X_k^{\mathrm{T}}]/\eta_A\|_F^2\\&=\max(S_{\lambda(\eta_A u_k)^{-1}}(A_k+(D-A_kX_k-E_k+\frac{Y_k}{u_k})X_k^{\mathrm{T}}/\eta_A),0)\end{aligned}\tag{5}$$

2. 固定 A,E,Y,u，更新 X：

$$\begin{aligned}X_{k+1}&=\arg\min_X \beta\|X\|_*+\frac{\eta_X u_k}{2}\|X-X_k+[-A_{k+1}^{\mathrm{T}}(D-A_{k+1}X_k-E_k+\frac{1}{u_k}Y_k)]/\eta_X\|_F^2\\&=D_{\beta(\eta_X u_k)^{-1}}(X_k+A_{k+1}^{\mathrm{T}}(D-A_{k+1}X_k-E_k+\frac{1}{u_k}Y_k)/\eta_X)\end{aligned}\tag{6}$$

令：$(U,\Sigma,V)=svd(X_k+A_{k+1}^{\mathrm{T}}(D-A_{k+1}X_k-E_k+\frac{1}{u_k}Y_k)/\eta_X)$，$svd(Mat)$ 定义为对矩阵 Mat 进行 svd 分解，则：

$$X_{k+1}=US_{(\eta_X u_k)}[\Sigma]V^{\mathrm{T}}\tag{7}$$

式中，$S_{(\eta_X u_k)}$ 为收缩算子，其定义同上文的 $S_\tau(x)$。

3. 固定 A,X,Y,u，更新 E：

$$\begin{aligned}E_{k+1}&=\operatorname{agr}\min_E\|E_k\|_1+\frac{u_k}{2}\|D-A_{k+1}X_{k+1}-E_k+\frac{1}{u_k}Y_k\|_F^2-\frac{1}{2u_k}\|Y_k\|_F^2\\&=\operatorname{agr}\min_E\|E_k\|_1+\frac{u_k}{2}\|E_k-(D-A_{k+1}X_{k+1}+\frac{1}{u_k}\end{aligned}$$

$$Y_k)\|_F^2-\frac{1}{2u_k}\|Y_k\|_F^2$$

$$=\max(S_{1/u_k}(D-A_{k+1}X_{k+1}+\frac{1}{u_k}Y_k),0)+\frac{1}{u_k}Y_k),0) \tag{8}$$

式中，$S_{(\beta/u_k)}$为一个收缩算子，收缩算子 $S_\tau(x)$定义如下：

$$S_\tau(X)=\begin{cases}x-\tau, & \text{如果 } X>\tau\\ x+\tau, & \text{如果 } X<-\tau\\ 0, & \text{其他}\end{cases} \tag{9}$$

4. 更新 Y：

$$Y_{k+1}=Y_k+u_k(D-A_{k+1}X_{k+1}-E_{k+1}) \tag{10}$$

5. 更新 u：

$$u_{k+1}=\min(u_{\max},\rho u_k) \tag{11}$$

式中，$u_{\max}$为 u 的上限，ρ 的定义如下：

$$\rho=\begin{cases}\rho_0, & u_k\max(\sqrt{\eta_A}\|A_{k+1}-A_k\|,\sqrt{\eta_X}\|X_{k+1}-X_k\|,\\ & \quad \|E_{k+1}-E_k\|)/\|D\|\leqslant\varepsilon_1\\ 1, & \text{其他}\end{cases} \tag{12}$$

式(4)的收敛标准有两个，同时满足这两个条件时算法收敛：

$$\begin{cases}(1)\|D-AX-E\|/\|D\|<\varepsilon_0\\ (2)u_k\max(\sqrt{\eta_A}\|A_{k+1}-A_k\|,\sqrt{\eta_X}\|X_{k+1}-X_k\|,\\ \quad \|E_{k+1}-E_k\|)/\|D\|\leqslant\varepsilon_1\end{cases} \tag{13}$$

式中，ε_0 和 ε_1 为设定的常数。

上述的求解过程可以总结为如下：

算法 1 STMLRC 模型求解算法

输入：词项-文档矩阵 D，主题空间维数 T，规范化参数 λ 和 β

初始化：设置 $\varepsilon_0>0,\varepsilon_1>0,\mu_{max}\gg u_0>0$，$k=0,A_0=(I_T,0),X_0=(I_N,0)$，最大循环次数 max Iter，$\eta_A>\|X\|^2$，$\eta_X>\|A\|^2$

求解：

While k≤max Iter 且 A 和 X 不满足收敛标准：

(1)通过式(5)—式(10)计算 A,X,E,Y.

(2)更新 μ.

End

输出：投影矩阵 A 和编码矩阵 X

在上述算法中，λ 和 β 分别控制投影矩阵的稀疏度和低秩约束在模型中的重要程度，λ 越大则投影矩阵的稀疏度越小；反之稀疏度越大。β 越大则低秩约束在模型中越重要；反之，重要性越低。根据实验的经验值，λ 和 β 的取值大小差距最好在一个数量级之内。

3 实验结果

3.1 实验数据集

本文选择 20NG[10] 和 colon-cancer[11] 作为实验数据集。其中，20NG 是目前常用的文本分析数据集，包括 20000 个文档，涉及 20 个不同的新闻组。在本文实验中，我们选择其中的两个新闻组“alt. atheism”和“talk. religion. misc”作为实验数据，并使用文档的 tf-idf 特征作为文档的特征向量。colon-cancer 数据库提供的是结肠癌细胞的数据。各个数据集中包含的数据样本情况如表 1 所列。

表 1 实验数据的统计

	20NG	colon-cancer
样本数	1427	62
特征维数规模	17390	2000
类别数	2	2

3.2 主题空间分类性能比较

为了验证 STMLR 所学习主题空间的分类性能，本文分别利用 STMLRC 和 LSA 构造主题空间，然后把数据映射到该主题空间，最后利用线性支持向量基学习分类器进行分类。在实验中，本文随机将数据集按照 2∶1 的比例分成训练数据集/测试数据集，然后使用 LibSVM[12] 工具包的线性 SVM 分类器对隐含主题空间的数据进行分类，采用 5-fold 交叉验证的方法在$[2^{-5},2^{10}]$区间内寻找分类器的最优规范化参数“c”。实验所基于的数据集是 20NG 数据集和 colon-cancer 数据集。实验结果如表 2 所列。

表 2 在不同维度下主题空间的分类性能

数据集	20NG		colon-cancer	
(算法)	LSA	STMLRC	LSA	STMLRC
10 维	85.96%	76.00%	100%	83.33%
100 维	94.04%	86.12%	100%	100%
500 维	94.38%	93.58%	100%	100%
1000 维	92.11%	95.66%	100%	100%
2000 维	93.99%	94.29%	100%	100%

从表 2 可以看出，在主题空间维度比较低的情况下，基于 STMLRC 算法得到的主题空间的分类性能将低于利用 LSA 算法得到的主题空间的分类性能。从信息论角度来看，对数据降维会导致信息丢失。为保证映射矩阵稀疏性，STMLRC 方法不可避免地比 LSA 方法丢失更多的信息，导致对应主题空间的表达能力略微下降，因此分类性能也将下降。随着空间维度增加，稀疏性导致的信息丢失也将越来越小，STMLRC 算法得到的主题空间分类性能也在不断提升。当主题空间达到一定维度后，STMLRC 算法的性能将超过经典的 LSA 算法。

3.3 主题空间映射矩阵稀疏性的比较

与稠密矩阵相比，稀疏矩阵的一个重要优势就是可以大幅减少主题空间映射矩阵的存储代价。为此，本文分别统计了 STMLRC 方法和 LSA 方法得到的主题空间映射矩阵的非零元素个数比，以此来考察主题空间映射矩阵的稀疏性。实验所基于的数据集是 20NG 数据集和 colon-cancer 数据集。具体结果如表 3 所列。

表 3 在不同数据集上主题空间映射矩阵稀疏性(百分比越小越好)

数据集	20NG		colon-cancer	
(算法)	LSA	STMLRC	LSA	STMLRC
10 维	100%	1.34%	100%	52.09%
100 维	100%	0.59%	100%	32.27%
500 维	100%	0.57%	100%	6.49%
1000 维	100%	0.59%	100%	3.27%
2000 维	100%	0.43%	100%	1.66%

从表 3 可以看出，STMLRC 方法得到的主题空间映射矩阵的稀疏性在绝大多数情况下远远优于 LSA 方法得到的映射矩阵。这也意味着，STMLRC 方法所得的到主题空间映射矩阵具有更小的存储代价。另外，随着主题空间维度增加，STMLRC 方法得到的主题空间映射矩阵稀疏性会降低。这种特性使得 STMLRC 方法更加适合于文本检索等实际应用。

3.4 低秩约束条件的影响

与 SLSA 方法相比，STMLRC 方法不仅考虑主题空间映射矩阵的稀疏性，也考虑了数据在主题空间中的分布特性（要求数据在主题空间中呈现出更好的聚类特性），因此所构造的主题空间具有更好的分类性能。为了考察低秩约束所带来的影响，本文在 colon-cancer 数据集上对比了 STMLRC 方法和 SLSA 方法所学习的主题空间的分类性能。实验结果如表 4 所列。

表 4 在不同数据集上主题空间的分类效果对比

数据集	20NG		colon-cancer	
（算法）	SLSA	STMLRC	SLSA	STMLRC
10 维	74.20%	76.00%	81.33%	83.33%
100 维	91.39%	86.12%	81.33%	100%
500 维	93.66%	93.58%	91.67%	100%
1000 维	92.61%	95.66%	91.83%	100%
2000 维	93.88%	94.29%	93.67%	100%

从表 4 可以看出，在相同空间维度下，STMLRC 方法所得到的主题空间中的分类效果总是优于 SparseLSA 方法。这也间接证明，低秩约束条件使得数据在主题空间具有更好的聚类特性，更有利于分类。同时，为了考察低秩约束条件对映射矩阵稀疏性的影响，本文分别统计了 STMLRC 方法和 SLSA 方法所得到的映射矩阵的非零元素百分比，结果如表 5 所列。从表中可以看出，低秩约束条件同时也可以降低映射矩阵的稀疏性，使得主题空间语义更加明确。

表 5 在不同数据集上映射矩阵的稀疏性对比

数据集	20NG		colon-cancer	
（算法）	SLSA	STMLRC	SLSA	STMLRC
10 维	1.4%	1.34%	97.11%	52.09%
100 维	0.56%	0.59%	95.70%	32.27%
500 维	0.31%	0.57%	91.80%	6.49%
1000 维	0.24%	0.59%	88.29%	3.27%
2000 维	0.14%	0.43%	83.40%	1.66%

结束语 本文针对传统的潜在语义分析模型存在的不足提出了低秩稀疏主题模型。该模型通过对映射矩阵施加稀疏约束来降低矩阵存储代价，使得主题语义更加明确；通过对编码系数矩阵施加低秩约束，使得数据在主题空间呈现出更好的聚类特性。实验结果也表明，基于本文所提出的稀疏主题模型得到的主题空间更有利于分类。但是，本文模型求解过程中涉及到矩阵的 SVD 分解，映射矩阵学习阶段仍需要较大的计算量。在后面的研究中，如何提高映射矩阵的计算速度是未来的一个重要研究方向。

参 考 文 献

[1] Dumais S T. Latent Semantic Analysis[J]. Annual Review of Information Science and Technology, 2005, 38(1): 188-230

[2] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407

[3] ChenX, Qi Y, Bai B, et al. Sparse Latent Semantic Analysis[C]// SIAM 2011 International Conference on Data Mining. 2011

[4] Liu G, Lin Z, Yu Y. Robust subspace segmentation by low-rank representation[C]// Proceedings of the 26th International Conference on Machine Learning. Haifa, Israel. Citeseer, 2010

[5] Liu Guang-can, Lin Zhou-chen, Yan Shui-cheng, et al. Robust Recovery of Subspace Structures by Low-Rank Representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 2233-2246

[6] Zhuang Lian-sheng, Gao Hao-yuan, Lin Zhou-chen, et al. Non-Negative Low Rank and Sparse Graph for Semi-Supervised Learning[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 2012

[7] Lin Z, Chen M, Wu L, et al. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices[R]. UIUC Technical Report UILU-ENG-09-2215. 2009

[8] Candès E. Compressive sampling[C]// Proceedings of the International Congressof Mathematicians. 2006

[9] Candès E, Li X, Ma Y, et al. Robust principal component analysis[J]. Journal of the ACM, 2011, 58(3)

[10] http://people.csail.mit.edu/jrennie/20Newsgroups/

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#rcv1.multiclass

[12] Chang C, Lin C. LIBSVM: a library for supportvector machines [OL]. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001

（上接第 79 页）

[6] Olguln D O, Pentland A S. Human activity recognition: Accuracy across common locations for wearable sensors, 2006[C]// Proceedings of International Symposium on Wearable Computers. 2006: 11-13

[7] Kunze K, Lukowicz P. Dealing with sensor displacement in motion-based onbody activity recognition systems[C]// Proceedings of the 10th international conference on ubiquitous computing. ACM, 2008: 20-29

[8] Forster K, Roggen D, Troster G. Unsupervised classifier self-calibration through repeated context occurences: is there robustness against sensor displacement to gain? [C]// International Symposium on Wearable Computers, 2009 (ISWC'09). IEEE, 2009: 77-84

[9] Lester J, Choudhury T, Borriello G. A practical approach to recognizing physical activities[M]// Pervasive Computing. Springer Berlin Heidelberg, 2006: 1-16

[10] Chavarriaga R, Bayati H, Millán J D. Unsupervised adaptation for acceleration-based activity recognition: robustness to sensor displacement and rotation[J]. Personal and Ubiquitous Computing, 2013, 17(3): 479-490

[11] Lonardi J L E K S, Patel P. Finding motifs in time series [C]// Proc. of the 2nd Workshop on Temporal Data Mining. 2002: 53-68

[12] Chiu B, Keogh E, Lonardi S. Probabilistic discovery of time series motifs[C]// Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2003: 493-498

[13] Lin J, Keogh E, Wei L, et al. Experiencing SAX: a novel symbolic representation of time series[J]. Data Mining and Knowledge Discovery, 2007, 15(2): 107-144