文章编号:1000－2367(2015)03－0155－06 DOI:10.16366/j.cnki.1000－2367.2015.03.030

SAX 结合 Adaboost 算法的时间序列分类问题

宋 玉,高明磊,宋 伟

(郑州大学 信息工程学院,郑州 450001)

摘 要:SAX 是一种典型的符号化特征表示方法.该方法在时间序列特征表示中不仅可以有效地降维、降噪,而且具有简单、直观等特点.时间序列长度不一、特征表示过程中信息损失等问题的存在,使得常规的分类算法难以很好地完成分类任务.在对时间序列数据进行基于 SAX 符号化的 BOP 表示方法的基础上,提出了结合集成学习中 AdaBoost 算法进行分类的新方法,实验结果表明,该方法不仅能很好地处理 SAX 符号化表示中的信息损失问题,而且与已有方法相比,在分类准确度方面也有了显著的提高.

关键词:时间序列;分类;SAX;BOP;AdaBoost

中图分类号:TP311 **文献标志码**:A

作为一种与时间变化相关的数据,时间序列通常表现出高维的特征,而且往往伴随着噪声的存在.时间序列数据是数据挖掘领域中重要的研究对象之一,广泛存在于科学研究、工程应用、金融服务、生物医疗等众多领域中.因此,对于大量的此类复杂数据,如何有效地挖掘和获取信息与知识,对科学研究以及生产实践都具有十分重要的价值和意义.

由于时间序列的高维特性及产生过程中出现的噪声信息,在数据挖掘和知识获取之前,通常需要对原始数据进行变换表示,将原始时间数据序列映射到新的低维特征空间中,以进行有效的数据降维并去除噪声,从而起到减少计算代价、提高数据挖掘与信息获取效率的作用.目前在该领域的研究过程中,已产生分段线性表示方法、符号化表示方法、基于域变换的表示方法等多种时间序列特征提取和表示方法,对时间序列数据挖掘及特征表示与相似性度量相关介绍可参考文献[1－2].

近年来,国内外研究者开始将符号化表示作为一种有效的离散化时间序列降维方法进行研究和关注,在这些研究中,由 Lin 和 Keogh 等人提出的基于分段累积近似法 (Piecewise Aggregate Approximation, PAA)的符号累积近似 (Symbolic Aggregate approximation,SAX)方法被认为是一种最为典型的符号化表示方法[3].SAX 方法提供了利用文本挖掘与生物信息的相关算法来解决时间序列数据挖掘中常见的分类、聚类、模式发现、异常检测和可视化等问题,该方法首先利用 PAA 方法将时间序列进行分段均值表示,之后将这些均值转化为离散化的字符表示,从而达到了降维降噪的目的,并且其符号化距离度量方法满足下界要求,SAX 方法在数据挖掘任务中的应用可参考文献[4－6]等.

将 SAX 方法应用于数据分类相关问题,目前的研究已取得了一定的成果[4,7],但由于 SAX 方法本身的信息损失问题,若有些损失的信息恰好是较为重要的特征信息,则对分类结果往往带来直接的影响,分类的准确性有待于进一步提高.针对此问题,本文提出了将 SAX 方法与集成学习中 AdaBoost 方法相结合,弥补信息损失从而提高时间序列数据分类准确率的方法,设计了相应实验并分析了实验结果.

实验过程中,首先利用 University of California Riverside 提供的时间序列分类聚类数据集,对比了使用基于 SAX 符号化的 BOP(Bag of Patterns,BOP)方法前后的分类表现,证明了 SAX 方法的有效性.同时,通过对实验结果的分析,发现在部分数据集上 BOP 方法却不如未采用 SAX 方法而对数据直接分类的表现,说

收稿日期:2014-10-23;**修回日期**:2015-03-10.

基金项目:国家自然科学基金(61202207);河南省教育厅科学技术研究重点项目(13A520453).

作者简介(通信作者):宋 玉(1969－),男,河南邓州人,郑州大学副教授,研究方向为数据挖掘及自然语言处理、人工智能等,E-mail:ieysong@zzu.edu.cn.

明该方法存在一定的不足,即:仅考虑分段序列的均值,有可能带来一定的信息损失,从而会影响到分类的准确性.为了克服由于信息损失影响分类准确性的问题,在实验中将 SAX 符号化 BOP 方法与集成学习方法相结合,采用 AdaBoost 算法[8]来整合各种不同参数下 BOP 表示后的信息多样性,以达到互相弥补信息缺失的效果[7],实验结果表明,分类的准确性得到了显著提高.

1 SAX 及 BOP 方法

对时间序列数据进行合适的特征表示,是进行相似性度量及数据挖掘的基础,将有助于保证和提高数据挖掘的效果.

1.1 SAX 方法

SAX 方法是 Lin 与 Keogh 等人在 PAA 基础上提出的时间序列符号化离散表示方法[3-4].该方法首先将时间序列转化为 PAA 表示,然后将其转换为符号化离散字符串.对于时间序列 $C=\{c_1,c_2,\cdots,c_n\}$,对其 SAX 表示的具体步骤如下:

1) 规格化.将时间序列 C 转换为均值为、标准差为 1 的标准序列 C'.

2) 降维.对 C' 进行 PAA 表示,得到 $\overline{C}=\{\overline{c_1},\overline{c_2},\cdots,\overline{c_i},\overline{c_w}\}$.$w$ 为时间序列 PAA 表示的分段数,

$$\overline{C_i}=\frac{w}{n}\sum_{j=\frac{n}{w}(i-1)+1}^{w}c_j. \tag{1}$$

其中 i 表示第 i 段,j 表示序列的第 j 个值点.

3) 离散化.根据选定的字母集大小,在高斯分布表中查找区间分裂点 β_i,将 PAA 表示映射为对应字符,最终得到离散化的字符串 $\hat{C}=\{\hat{c}_1,\hat{c}_2,\cdots\hat{c}_w\}$,

本文以实验中采用的 ECG 数据集为例,给出了 SAX 符号化表示的可视化样例.

SAX 方法简单易用,能够有效降维以及进行相似性度量,且满足下边界引理,在时间序列的分类、聚类、模式识别与发现以及可视化中具有良好的性能.

1.2 BOP 方法

相似性度量方法是时间序列数据挖掘中分类、聚类、异常检测等任务的基础,不同的时间序列的表示有其适应的相似性度量方法.目前多数基于形态的相似性度量方法有较好的效果,但是随着要处理的时间序列的长度增加,对于较长的数据序列的处理效果表现不佳.欧式距离、动态时间弯曲(Dynamic Time Warping)[9]、编辑距离(Edit Distance)[10]、最大公共子串(Longest common subsequence)[11]等是较为常用的方法,但是以上这些方法对于不等长时间序列的度量、较长的时间序列的效率问题,以及序列中部分长度序列值本地相关等数据没有统一较好的解决方案.

BOP 方法是 Lin 等人在 SAX 方法的基础上,借鉴了自然语言处理中 BOW(Bag of Words)的表示方法,利用直方图方法来对时间序列进行相似性度量.自然语言处理中一个常见问题是发现文档间的相似性,BOW 方法将文档表示为无序单词的组合,利用相同单词在不同文档中的出现频率作为度量文本相似性的依据.在 VSM(Vector Space Model)模型中[12],可以将 m 个文档中的每个文档用一个向量 $D_i=(d_{i1},d_{i2},\cdots d_{in})$表示,并据此建立一个 $m\times n$ 矩阵 D,其中 m 为文档个数,n 为单词个数,d_{ij} 表示第 j 个单词在第 i 个文档中出现的频率.则文档相似度可以通过向量空间计算得到.BOP 方法的出现,使得解决上述相似性度量方法中所有的问题成为可能.

Lin 等人在文献[13]中对 BOP 相关算法及应用进行了描述,该方法的主要步骤为:

1)利用长度为 n 的滑动窗口对长度为 m 的时间序列提取子序列(其中 $n\ll m$),将每个子序列规格化为均值为,标准差为 1,并转换为 SAX 单词(Pattern),最终得到一系列字符串;

2)构建 BOP 矩阵 M.其中每一列对应一个时间序列,每一行对应一个 SAX 单词(Pattern),每个元素 M_{ij} 表示单词 i 在时间序列 j 中出现的频率;

3)采用类似自然语言处理中 BOW 的方法,进行相似性度量.

采用 BOP 方法,不同时间子序列间拥有相同或者相似结构的可以被标记为相似,而忽略了时间序列的

长度与出现的位置. 该方法对聚类、分类、异常检测等任务表现出了较高的性能.

2　AdaBoost 算法

集成学习(Ensemble Learning)是机器学习方法中较新的一种技术，是机器学习的热门方向之一. 目前主要的集成学习方法有 Bagging 算法、Boosting 算法和 Stacking 算法等，该方法使用一系列学习器进行学习，并使用某种规则把学习结果进行整合，从而获得比单个学习器更好的学习效果. 对于分类问题，集成学习在对新的实例进行分类的时候，把若干个单个分类器集成起来，通过对多个分类器的分类结果进行某种组合来决定最终的分类，以取得比单个分类器更好的性能.

AdaBoost 算法是目前使用最广泛的 Boosting 算法，该算法通过顺序给训练集中的数据项改变权重来创造不同的基础学习器，其核心思想是重复应用一个基础学习器来修改训练数据集，从而可以在预定数量的迭代下产生一系列基础学习器. 在训练之初，所有的数据项都被初始化为相同的权重，之后每次增强的迭代都会生成一个适应加权之后的训练数据集的基础学习器. 每一次迭代的错误率都会被计算，正确划分的数据项的权重被降低，错误划分的数据项权重将增大. Boosting 算法的最终模型是一系列基础学习器的线性组合，而且系数依赖于各个基础学习器的表现.

3　实验设计及结果评估

3.1　实验数据

实验数据选自 University of California Riverside 时间序列分类聚类数据集[14]. 为了测试 BOP 方法与 AdaBoost 算法的有效性，实验选取了时间序列波形与周期性具有代表性的 4 种二分类数据集进行测试，所有数据正负样本数相同以使得学习样本没有偏差. Yoga 数据波形平稳并呈现出一定的周期性，Lighting2 数据震荡剧烈又不失整体的趋势，是夹杂着噪声的季节性数据；ECG 数据为典型的生理医学数据代表；Coffee 波形震荡不如 Lighting2 数据剧烈，但是局部极值较多. 实验数据代表的时间序列具有较为广泛的代表性.

3.2　SAX 及 BOP 表示方法

采用基于 SAX 的 BOP 方法，可以对时间序列数据进行有效降维及相似性度量. 对时间序列数据进行 BOP 表示，需要设定参数簇(n,w,a)，3 个参数分别表示窗口长度 n，窗口内符号串(单词)长度 w 和字母表大小 a. 不同的参数簇设定将产生不同的 SAX 数据表示：从数据中提取的每个单词(pattern)对应的范围随着窗口长度 n 的增大而增加，较大的 n 值会产生低解析度表示，较小的 n 会产生高解析度表示，而改变 w 值对解析度的影响效果则相反；较大字母表 a 会将窗口内的值映射到更多更细的区间，而较小的字母表则会提高对于噪声和异常值的容忍度. 本实验中，对于时间长度为 L 的时间序列 T，参数簇在 $n=[0.10L, 0.11L, \cdots, 0.3L]$，$w=\{2, 4, 8, 16\}$，$a=\{3, 4, \cdots, 10\}$的范围中遍历.

对于长度为 L 的时间序列 T，每滑动一个窗口，对窗口内的时间序列进行 SAX 符号化后建立一个长度为 n 的 BOP. 如果滑动步长 t 为 1，则滑动窗口一共可以建立 $L-n+1$ 个子序列，对应就是 $L-n+1$ 个 BOP. 于是整个时间序列相似度可以表示为 BOP 间任意范数的距离或者 cosine 相似度等，这样不同时间子序列间拥有相同或者相似结构的可以被标记为相似，从而可以忽略时间序列的长度与出现的位置. 本实验中，采用最简单的方法，即通过统计不同的 BOP 表示出现的次数来衡量不同时间序列的相似度.

每一组(n,w,a)参数对应产生一种 BOP 符号化序列 因为 1-NN 分类器具有无参并且易于比较的优点，按照文献[14]中的建议，将所有符号化序列最终用 1-近邻(1-Nearest-Neighbor，1-NN)方法进行分类，并运用留一交叉验证法(Leaving-One-Out Cross Validation，LOOCV)方法防止过拟合. 实验采用 Python 语言进行相关算法实现，1-NN 分类采用了 WEKA 平台[15]. 在 4 种实验数据集上的 LOOCV 错误率如图 1 所示，其中纵坐标为错误率，横坐标为按照错误率排列的 BOP 参数升序表示.

对图 1 分析可以看出，基于 BOP 表示的 1-NN 分类器在 Coffee 数据集上表现优异，可以达到 100% 的准确率，在 lighting2 上的总体表现为 4 个数据集中最不理想的，但是最低错误率依然可以达到 16.5%，Yoga 数据集中虽然最低的错误率为 17%，但是整体参数的表现非常优异，不会因为参数的微小扰动而造成分

类结果的大幅度改变. 全部 BOP 表示的分类错误率均在 50%以下,意味着 SAX 符号化表示和 BOP 方法抓住了时间序列的统计特征. 表 1 列出了 BOP 表示方法在 1-NN 分类器下的最低错误率.

表 1 BOP 表示方法在 1-NN 分类器下的最低错误率

数据集	最低错误率	数据集规模	序列长度
EC	0.13	20	96
Lighting	0.17	12	63
yoga	0.17	30	42
Coffee	0.00	56	28

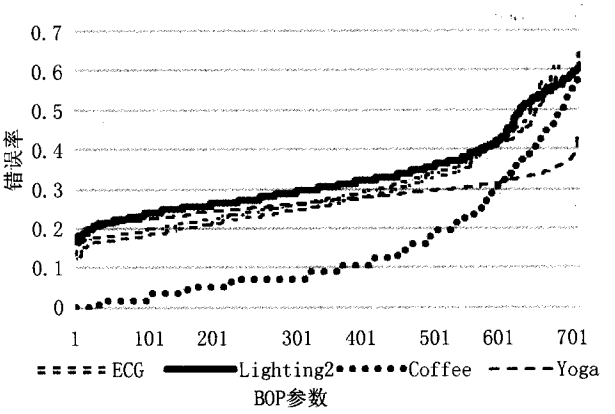

图1 采用BOP方法在实验数据集上的LOOCV错误率升序排序

为了考查参数设定对 BOP 特征提取的影响,表 2 列举了 LOOCV 错误率最低的 10 种 BOP 表示参数. 如上文提到的,一种 BOP 表示对应一组(n, w, a)参数簇. 表格最底部一行为 n,w 和 a 的平均值. 从表格中可以看出,对本文实验数据,移动窗口的长度 n 在 0.22 到 0.28 的平均值之间表现较好,窗口内字符数 w 没有显著的统计规律,而字母表长度 a 在 7 或者 8 的解析度更能在不失去噪声与异常点鲁棒性的情况下抓住时间序列的统计特征. 7 或者 8 在测试的字母表顺序中属于较大的参数取值,如前文讨论,大的字母表顺序可以勾画出时间序列更细节的特征,但是会失去部分对参数和异常点的鲁棒性. 在标准数据集上的测试因为数据本身的噪声不大,因此高的解析度可以更好的提取特征,然而如果需要处理更高噪声的时间序列,字母表则不宜太大,否则会受到噪声和异常值较多的干扰. 为了抓取更好的特征,时间序列的预处理及参数的选择将会是重要的解决方法,此问题会在以后的研究中进行更加深入的探讨.

表 2 LOOCV 错误率最低的 10 种 BOP 参数统计

ECG			Lighting2			Coffee			Yoga		
n	w	a	n	w	a	n	w	a	n	w	a
0.31	4	7	0.33	2	7	0.20	4	10	0.22	4	10
0.29	4	7	0.23	4	6	0.19	16	6	0.22	8	9
0.32	4	7	0.27	2	7	0.16	4	10	0.26	4	9
0.16	2	10	0.30	2	7	0.24	16	6	0.21	8	9
0.30	4	7	0.30	2	8	0.26	16	6	0.24	4	8
0.19	2	5	0.17	4	7	0.17	16	6	0.16	2	10
0.33	4	8	0.18	4	7	0.23	4	10	0.23	4	10
0.33	4	10	0.15	4	7	0.24	8	8	0.24	4	6
0.35	2	10	0.16	4	7	0.19	8	6	0.23	4	9
0.23	4	8	0.36	2	10	0.31	8	5	0.22	4	9
0.28	3.4	7.9	0.24	3	7.3	0.22	10	7.3	0.22	4.6	8.9

为了对比 SAX 符号化方法与不使用符号化方法的分类表现,表 3 截取了由 Tony Bagnall 测试的 UCR 时间序列数据集在 weka 上的分类表现[14],该结果展示了以上 4 种二分类数据集未使用 BOP 表示处理的错误率. 分类器分别选取了 1-最近邻算法(1-NN),贝叶斯分类器(NB),决策树(C4.5),多层感知机(MLP),Random Forest(RF)算法,逻辑模型树(LMT) 和支持向量机(SVM),最右列为 7 种分类器在 4 个实验数据集上最低的错误率.

表 3 4 种数据集未使用 BOP 符号化表示的错误率

数据集	Knn	NB	C4.5	MLP	RF	LMT	SVM	最小错误率
ECG	0.11	0.23	0.28	0.16	0.19	0.18	0.19	0.11
Lighting2	0.20	0.33	0.38	0.26	0.21	0.36	0.28	0.20
Yoga	0.17	0.46	0.30	0.26	0.22	0.28	0.37	0.17
Coffee	0.25	0.32	0.43	0.04	0.25	0.00	0.04	0.00

对比表 1 与表 3 可以看出,使用 BOP 表示后,ECG 最低错误率有所增高,而在其他 3 个数据集上的 1-

NN 表现均不高于 7 种分类器中的最低错误率. 因此可以得出结论，SAX 符号化表示方法后的 BoP 能够较好的抓住时间序列的统计特征，但对于某些数据集效果不明显.

3.3 AdaBoost 方法

采用 BOP 表示，难免存在信息损失，比如对于 ECG 数据，运用 BOP 表示后的表现甚至不如直接使用原始数据进行分类的结果. 为解决该问题，本文运用集成学习方法来整合多种参数下的 BOP 表示后的信息多样性，以克服单个 BOP 符号化表示后的信息缺失导致的分类精度下降问题，从而达到互相弥补信息缺失的效果. 实验采用 AdaBoost 算法，对实验数据集进行训练分类，AdaBoost 算法通过不断迭代分类器，根据分类效果改变训练样本的权重，将难分的样本（SAX 符号化后信息损失过大的样本）着重进行训练分类，从而达到以提高分类器性能来弥补信息损失的效果. 表 4 列出了前 20 轮迭代直到收敛后的错误率. 对于前述实验中错误率已达到 0%的 Coffee 数据集之外的其他 3 个数据集，前 50 轮 AdaBoost 迭代后的错误率收敛曲线如图 2 所示.

表 4　基于 SAX+BOP 的 AdaBoost 方法前 20 次迭代的分类错误率

迭代次数	ECG	Lighting2	Yoga	Coffee	迭代次数	ECG	Lighting2	Yoga	Coffee
1	0.13	0.17	0.17	0.00	11	0.09	0.02	0.13	0.00
2	0.13	0.17	0.17	0.00	12	0.10	0.03	0.13	0.00
3	0.10	0.09	0.17	0.00	13	0.09	0.02	0.13	0.00
4	0.12	0.11	0.17	0.00	14	0.09	0.02	0.12	0.00
5	0.11	0.07	0.17	0.00	15	0.09	0.00	0.12	0.00
6	0.12	0.06	0.17	0.00	16	0.09	0.02	0.11	0.00
7	0.10	0.03	0.17	0.00	17	0.09	0.01	0.11	0.00
8	0.10	0.05	0.17	0.00	18	0.09	0.01	0.11	0.00
9	0.10	0.02	0.17	0.00	19	0.09	0.00	0.08	0.00
10	0.10	0.03	0.17	0.00	20	0.09	0.01	0.09	0.00

对图 2 进行分析，经过 AdaBoost 训练后的弱 BOP 表示，能够根据分类结果的权重来加权投票的效果，从而弥补了一些非常鲁棒的信息损失样本造成的误差. 在 3 个数据集上的迭代次数在 20 次左右均可以收敛，不会因为迭代次数过多而造成过拟合现象[16]. 因此，SAX 符号化结合集成学习 AdaBoost 算法，在时间序列的表示与分类问题上有着优异的表现.

表 5　各种方法最优分类结果

	7 Classifier	SAX+O	SAX+OP+AdaBoost
EC	0.11	0.13	0.08
Lighting	0.2	0.17	0
Yoga	0.17	0.17	0.07
Coffee	0	0	0

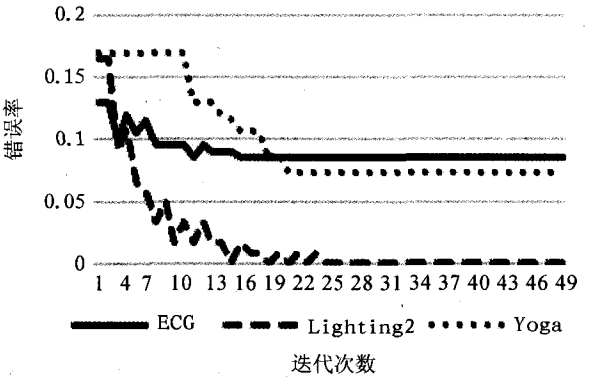

图2 AdaBoost错误率收敛曲线

4　结束语

基于 SAX 的 BOP 方法对异常值和噪声具有良好的鲁棒性，能够对时间序列数据有效降维，本文提出并通过实验验证了 BOP 方法与集成学习方法相结合对于二分类问题的有效性，大大降低了分类的错误率. 表 5 列出了本文中实验方法与 7 种直接分类方法在实验数据集上的结果对比. 从实验结果不难看出，BOP 方法与集成学习 AdaBoost 算法的结合，可以有效弥补 SAX 符号化表示后的信息缺失，提高分类的准确率. 结合本文的研究结果，下一步工作将对 BOP 方法中参数的选取对分类结果的影响，以及多分类问题进行研究.

参　考　文　献

[1] Fu T. A review on time series data mining[J]. Engineering Applications of Artificial Intelligence, 2011, 24(1): 164-181.

[2] Li Hai-lin, Guo Chong-hui. Survey of feature representations and similarity measurements in time series data mining[J]. Application Re-

search of Computers,2013,30(5):1285-1291.

[3] Lin J,Keogh E,Lonardi S,et al. A symbolic representation of time series, with implications for streaming algorithms[C]. Proc of the 8th ACM SIGMOD Workshop on Research issues in data mining and knowledge discovery(DMKD '03),San Diego,2003.

[4] 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.

[5] Junejo I, Aghbari Z. Using SAX representation for human action recognition[J]. Journal of Visual Communication and Image Representation,2012,23(6): 853-861.

[6] Afroni M, Sutanto D,Stirling D. Analysis of Nonstationary Power-Quality Waveforms Using Iterative Hilbert Huang Transform and SAX Algorithm[J]. IEEE Transactions on Power Delivery, 2013,28(4): 2134-2144.

[7] Oates T,Mackenzie C,Stein D,et al. Exploiting Representational Diversity for Time Series Classification[C]. Proc of 11th Int Conf on Machine Learning and Applications(ICMLA '12), Boca Raton, 2012.

[8] Freund Y,Schapire R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting[J]. Journal of Computer and System Sciences,1997, 55(1):119-139.

[9] Keogh E, Pazzani M. Derivative dynamic time warping[C]. Proc of 1st Int Conf on Data Mining, Chicago,2001.

[10] Chen L,Ng R. On the marriage of lp-norms and edit distance[C]. Proc of 30th Int Conf on Very Large Data Bases(VLDB '04), Morgan Kaufmann,2004.

[11] Bergroth L,Hakonen H,Raita T. A survey of longest common subsequence algorithms[C]. Proc of 7th Int Symp on String Processing and Information Retrieval(SPIRE 2000), Coruña,2000.

[12] Salton G,Wong A,Yang C. A Vector Space Model for Automatic Indexing[J]. Communications of the ACM, 1975,18(11):613-620.

[13] Lin J, Khade R, Li Y. Rotation-invariant similarity in time series using bag-of-patterns representation[J]. Journal of Intelligent Information Systems,2012,39(2):287-315.

[14] Keogh E, Zhu Q,Hu, B,et al. The UCR Time Series Classification/Clustering Homepage [EB/OL]. [2014-03-10]. http://www.cs.ucr.edu/eamonn/time_series_data/.

[15] Hall M, Frank E,Holmes G. The WEKA data mining software: an update[J]. ACM SIGKDD Explorations, 2009,11(1):10-18.

[16] Schaffer C. Overfitting avoidance as bias[J]. Machine Learning,1993,10(2):153-178.

Research on Time Series Data Classification Combine SAX and AdaBoost Algorithm

SONG Yu, GAO Minglei, SONG Wei

(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: Symbolic Aggregate approXimation (SAX) is a typical symbolic representation method, which is straight-forward and very simple, and it efficiently converts time series data to a symbolic representation with dimension reduction. The issues of time series data such as variable in length, and information lose during the representation, making many traditional classification methods unable to apply directly. This paper focus on the SAX discretization method coupled with the Bag of Patterns (BOP) representation in classification task, and proposed the new approach by use AdaBoost Algorithm to remedy the information loss by SAX representation. The experimental results show that, the approach improved the classification accuracy obviously.

Keywords: time series; classification; SAX; BOP; AdaBoost