一种基于滑动窗口模式匹配的加权预测方法

王丽珍 周丽华 邓世昆
(云南大学滇池学院理工学院计算机科学与工程系 昆明 650091)

摘 要 随着中国改革开放的不断深入和社会经济的持续发展,各种社会矛盾逐渐复杂化和多样化,社会治安面临空前的挑战。基于社会治安情况的历史数据,对未来一段时期内的治安状况做出科学的预测,将使治安管理工作事半功倍。数据挖掘是指从大量数据中挖掘出有趣的模式和规则,并根据挖掘结果做出科学的判断或预测的技术。目前,在社会治安状态预测方面的研究报道还很少,预测结果的准确率也始终困扰着我们,研究一种新颖的、高准确率的预测方法是我们的共同期待。据此,提出一种基于滑动窗口模式匹配的加权预测方法,大量的实验以及实际应用的结果表明,该算法具有简单、稳定、高准确率等特点。

关键词 滑动窗口,模式匹配,加权
中图法分类号 TP311 **文献标识码** A

Weighted Prediction Method Based on Sliding Window and Pattern Matching

WANG Li-zhen ZHOU Li-hua DENG Shi-kun
(School of Computer Science and Engineering, Dianchi College of Yunnan University, Kunming 650091, China)

Abstract With the deepening of Chinese reform and opening to the outside world, and developing sustainably of the society and economy, various social conflicts become complex and diverse. As a result, public security is facing unprecedented challenge. At this time, if we could make a scientific prediction on the future social stability based on the historical data of the public security, our public security management work would get two fold results with half the effort. Data mining refers to extracting or discovering interesting data patterns or rules hidden in large data sets, and makes a scientific judgment or prediction according to these discovered patterns or rules. So far, research on social stability early warning is very rare, and the accuracy of prediction results is always a difficult problem. In this paper, a novel and high accurate prediction method based on sliding window and pattern matching was proposed. Extensive experiments and the actual applications show that the proposed algorithm has features of simplicity, stability and high accuracy.

Keywords Sliding windows, Pattern matching, Weighted

1 引言

当前,中国正处在社会经济持续较快发展的黄金时期。多年来,对旧体制的革新与更替,以及对新体制的尝试与创新,其广度和深度可谓前所未有。也正因为此,各种社会矛盾也空前复杂和多样,导致的社会治安事件也越来越多,处理起来也越来越复杂,付出的社会经济成本也越来越高。因此,建立一个有效的社会治安状况预测系统,对未来一段时间内的治安状况做出较为准确的预测,从而科学合理地部署警力或提前干预,将对社会经济的持续、稳定与发展提供有力的保证。

数据挖掘旨在从大量的历史数据中挖掘出有价值的模式与规则,对未来事态发展做出科学的预测。数据挖掘技术发展到今天,已经产生了一些具备实用价值的预测算法。如传统的回归统计技术[1,2],通过一个或一组变量的取值估计(预测)另一个变量的取值。如果两个变量在散点图上基本呈线性关系,则可用一元线性回归方程来描述。而在许多实际应用中,影响预测指标的因素常常不只一个,例如影响害虫盛发期的生态因素有温度、湿度、雨量等,这时需要应用多元线性回归技术。此外,有的问题还需要使用非线性回归技术才能解决。当然,回归统计技术解决的是明确了影响因素的预测问题。对于更为广泛的应用领域,学者们提出了更多的方法,如决策树[3,4]、人工神经网络[1,2]、贝叶斯分类器[5]、支持向量机[6,7]等,均取得了一定的应用效果。其中,基于支持向量机及其改进的算法被应用于脸谱识别[8]、手写体识别[9]、文本及图像分类[10]中。文献[11]针对未来潜在的客户流失进行预测。关于挖掘稀有类和不平衡数据的工作,文献[12]将多个分类器组合起来,让每一个分类模型都在其优势空间区域中

本文受国家自然科学基金(61472346,61262069),云南省自然科学基金(2015FB149,2015FB114)资助。
王丽珍(1962-),女,博士,教授,博士生导师,主要研究方向为数据库、数据挖掘及计算机算法等;周丽华(1968-),女,博士,教授,主要研究方向为社会网、决策支持等;邓世昆(1955-),男,教授,主要研究方向为互联网与网络安全等。

发挥作用。针对时间序列数据的预测研究，有电力系统时段负荷预测[13]、股市行情预测[14]、机场噪声预测[15]和时序关系预测[16]等。

总之，基于数据挖掘技术的各种预测算法已经广泛地应用到实际生活中，为社会经济发展做出了贡献。但是，有关社会稳定预警方面的算法，到目前为止，还基本停留在尝试阶段。为此，本文提出了一种基于滑动窗口模式匹配的加权预测算法，用于预测未来某一时期的社会治安状况。众所周知，很难有一种预测算法能够做到面面俱到，为此，进一步提出了一种加强算法，用以修正加权预测算法预测结果中不够理想的部分。大量的实验和实际应用验证了新算法在准确率、稳定性等方面的明显优势。

2 问题的定义与预测任务

2.1 问题的定义

面对多年积累的社会治安状况指标数据记录，如何从这些数据中挖掘出该地区的社会治安状况指标数据的变化趋势和规律，从而预测未来一段时期内，比如一个月、一个季度甚至半年，该地区的社会治安状况，以便提前做好矛盾化解、警力部署和其它预防措施，一直是社会治安主管部门的共同期盼。

例如，现有关于某地区过去若干年的治安状况指标数据记录，这些数据记录中包含了10个至数10个不同的指标，每个指标在每一个月都有一个相关记录。如表1所列，某地前3个治安状况指标分别是刑事案件立案数、刑事案件破案数和盗窃案件立案数。各指标的数据为2008年01月至2015年1月的相关历史记录。

表1 社会治安状况指标数据示例

时间	刑事案件立案数	刑事案件破案数	盗窃案件立案数	…
2008.01	93	12	68	…
2008.02	81	8	58	…
2008.03	89	10	64	…
2008.04	107	24	73	…
2008.05	86	65	55	…
2008.06	51	43	43	…
…	…	…	…	…
2015.02	预测	预测	预测	…
2015.03	预测	预测	预测	…

2.2 预测任务

现在要求基于表1所列的历史数据来预测未来该地区的社会治安状况指标数据。比如预测该地区第1个指标在未来4个月的值分别为48，36，126和98，其他指标也作类似预测处理，这样该地区的主管部门即可根据相关指标的预测值对未来4个月的社会治安工作做出部署，如针对一些可能的严重违法犯罪事件和突发性事件，做到有效的布控与处理，甚至做到防患于未然。这对有效利用社会资源、节省社会经济成本、维护人民群众生命财产安全和社会稳定等，都具有不可估量的意义。

2.3 传统的加权移动平均数法

当面对的数据是一个随时间变化的序列数据即时间序列时，采用自变量为时间 t 的回归分析方法对预测问题进行建模即 $y=f(t)$，然后采取趋势外推原理对未来数值进行预测，这就是所谓的时间序列预测法[17]。

可发现针对社会治安状况指标预测任务，时间序列预测法中的加权移动平均数法与多元回归的最小二乘法相比，效果明显。因此，本文将加权移动平均数法作为基础方法与新提出的算法进行比较。

3 相关定义及预测算法

定义1(滑动窗口模式与目标窗口模式) 指标 s 的历史数据序列$\{s[0],\cdots,s[n-1]\}$中，将长度为 L 的数据序列 moveGAP$[i,L]=\{s[i],\cdots,s[i+L-1]\}(i=0,1,\cdots,n-L)$称为该指标数据的 L 滑动窗口模式，而将时间上距离预测值最近的一个滑动窗口模式 moveGAP$[n-L,L]$称为目标窗口模式，记作 des_GAP$[n-L,L]$或 des_GAP。

例1 在表1中，如果定义滑动窗口模式的长度 $L=5$，则对应指标“刑事案件破案数”的滑动窗口模式有 moveGAP$[0,5]=\{12,8,10,24,65\}$，moveGAP$[1,5]=\{8,10,24,65,43\}$等。2015年1月(预测值)之前的滑动窗口模式是目标窗口模式 des_GAP。

3.1 基于滑动窗口模式匹配的加权预测算法

假设指标 s 的时间序列数据记录为$\{s[0],\cdots,s[n-1]\}$，要预测该指标未来第 m 个月的数据。设滑动窗口的长度为 $L=5$，于是，指标 s 的目标窗口模式 des_GAP 是 s 最近5个月的数据序列$\{s[n-5],\cdots,s[n-1]\}$。基于对大量实际数据序列的分析，提出如下的预测思想：1)在滑动窗口中，寻找与目标窗口模式相似的窗口模式；2)基于相似的窗口模式遴选与预测数据“相关”的历史数据；3)使用加权平均法求出指标 s 的预测值。

详述如下：从第1个滑动窗口 moveGAP$[0,5]=\{s[0],\cdots,s[4]\}$开始，考查 moveGAP$[i,5](i=0,1,\cdots,n-6)$与 des_GAP 的相似度，当相似度满足一定条件时(将在3.2节进行阐述)，$s[i+4+m]$(假设要预测未来第 m 个月的数据)为一个与预测数据“相关”的历史数据。之后，窗口向前(右)滑动1格(即1个月)，遴选第2个与预测数据“相关”的历史数据，直至最后一个滑动窗口 moveGAP$[n-6,5]$。

moveGAP$[i,L]$与 des_GAP 之间的相似度作为被遴选历史数据的权值。如果遴选结果集非空，则基于遴选的结果集及相应的权值，使用加权平均法求出指标 s 在未来第 m 个月的预测值。然后将滑动窗口长度 L 缩小一格，再重复上述过程，直到 $L=1$ 为止。

最后，以 L 的取值大小为权重，对以上各个不同 L 取值的相应预测结果求加权平均值并将其作为指标 s 在未来第 m 个月的最终预测值。

算法1描述了上述过程。

算法1 基于滑动窗口模式匹配的加权预测算法

```
1. for (每一个预测目标)/* 如预测未来第3个月的数据 */
2.   { for (每一个指标)
3.     {for(g=L;g>=1;g--)   /* 窗口大小从L到1 */
4.       {得到对应指标数据目标窗口 des_Gap;
5.         寻找与 des_Gap 相似的窗口模式;
6.         基于相似的窗口模式遴选与预测数据“相关”的历史数据;
```

7. 基于“相关”的历史数据计算目标数据；
}；
8. 根据不同 L 值计算所得的不同目标数据来确定最后的目标数据；
}；
}；

以上算法能顺利地将各个相关指标未来 m 个月的数据预测出来。

算法的时间复杂度分析：基于滑动窗口模式匹配的加权预测算法采用了 4 层循环结构，最外层执行了 *futureMonthCount*(预测目标数)次，次外层执行了 *caseRowCount*(预测指标数)次，第三层最多执行 L(滑动窗口长度)次，而最内层执行 *monthColCount*(滑动窗口数)次，故算法的时间复杂度不超过 O(*futureMonthCount*×*caseRowCount*×L×*monthColCount*)。

3.2 滑动窗口模式间的相似度

为了衡量滑动窗口模式与目标窗口模式之间的相似度，采用了经典的修正余弦相似度度量的方法[18]，因为一个窗口模式可以看成是一个多维向量，而余弦相似度实际是两个向量之间夹角(余弦)的度量。

对于某指标的窗口模式 t_a 和 t_b，记窗口中时间序列数据分别为$\{d_{a_i}, i=1,\cdots,L\}$和$\{d_{b_i}, i=1,\cdots,L\}$，分别计算它们的平均值$\overline{d_a}$ 和$\overline{d_b}$。于是，相似度计算公式见式(1)。

$$Sim(t_a,t_b)=\frac{\sum_{i=1}^{L}(d_{a\_i}-\overline{d_a})(d_{b\_i}-\overline{d_b})}{\sqrt{\sum_{i=1}^{L}(d_{a\_i}-\overline{d_a})^2}\sqrt{\sum_{i=1}^{L}(d_{b\_i}-\overline{d_b})^2}} \tag{1}$$

其中，$\sum_{i=1}^{L}(d_{a_i}-\overline{d_a})(d_{b_i}-\overline{d_b})$度量两窗口模式变化方向的统一程度，其值越大，说明两模式越趋向相同的方向变化；相反，说明两模式越趋向相反的方向变化。$\sqrt{\sum_{i=1}^{L}(d_{a_i}-\overline{d_a})^2}$ 和 $\sqrt{\sum_{i=1}^{L}(d_{b\_i}-\overline{d_b})^2}$ 度量两模式数据各自的离散程度，$\sqrt{\sum_{i=1}^{L}(d_{a\_i}-\overline{d_a})^2}\sqrt{\sum_{i=1}^{L}(d_{b\_i}-\overline{d_b})^2}$ 越大，模式的数据越离散，最终的相似度越小；相反，模式的数据越收敛，最终的相似度越大。

3.3 预测准确率的评估

本节将讨论如何评估预测的准确率。首先，第 i 次预测的准确率 Acc_i 用式(2)计算。

$$Acc_i=\begin{cases}1-\left|\dfrac{y'-y}{y}\right|, & \text{其值}\geqslant 0\\ 0, & \text{否则}\end{cases} \tag{2}$$

其中，y 代表原值，y'为预测所得值。

其次，将各个指标的历史数据一分为二，前者作为训练集，后者为测试集，训练集和测试集的比例可以分别取 5∶5，6∶4，…，9∶1。当取其比例为 p∶q 时，有 n 个月的历史数据，预测未来 m 个月的指标，这时，训练集的大小为$\left\lceil\frac{n\times p}{p+q}\right\rceil+m\times i$，其中，$i=1,\cdots,k$，而 $k=\max\{i\mid(\left\lceil\frac{n\times p}{p+q}\right\rceil+m\times i)\leqslant n\}$。将算法 1 中的历史数据改为动态变化的训练集和测试集数据，并统计每次预测结果的准确率。

最后，计算 k 次预测的平均准确率。

3.4 加强算法

算法 1 没有考虑各个预测指标之间的相互影响，但事实上，影响社会稳定的各指标数据之间有一定的相关性。因此，为了进一步提高预测结果的准确率，基于算法 1，进一步提出了一个加强算法 strongWarningBroadcast。

式(1)可以计算 2 个指标之间的相似度，同时如 3.3 节所述，可评估各指标预测结果的准确率。于是，如果满足如下两个条件和条件则调用加强算法 strongWarningBroadcast 修正算法 1 得到的预测结果。

条件 1　存在一个预测准确率不够满意的指标 S_x，同时存在一个预测准确率足以令人满意的指标集$\{S_{y,1},\cdots,S_{y,t}\}$($t\geqslant 1$)，这是执行算法 2 的必备条件之一；

条件 2　指标数据 S_x 至少与$\{S_{y,1},\cdots,S_{y,t}\}$($t\geqslant 1$)中一个指标 $S_{y,i}$($1\leqslant i\leqslant t$)间的相似度 $Sim[S_x, S_{y,i}]$达到一定的阈值(用户给定)，这是能够执行算法 2 的必备条件之二。

对于任何满足条件 1 和条件 2 的指标 S_x 和$\{S_{y,1},\cdots,S_{y,t}\}$，以 S_x 与 $S_{y,i}$($1\leqslant i\leqslant t$)的相似度值作为权值，求$\{S_{y,1},\cdots,S_{y,t}\}$($t\geqslant 1$)中各个指标预测结果的加权平均数，并以这个加权平均数修正指标 S_x 的预测值。

4 系统实现与分析

本节先介绍一个使用了滑动窗口模式匹配加权预测算法的预警系统，该系统已经被成功部署到现实应用中。然后，就滑动窗口模式匹配加权预测算法与传统的加权移动平均数法进行比较，以验证滑动窗口模式匹配加权预测算法的性能。

4.1 预警系统实现

预警系统采用 Visual studio 2008 C＃编写，后台数据库采用 SQL Server 2008。下面以一个用户角色，按系统的使用流程对其进行介绍。

4.1.1 主界面

如图 1 所示，主界面共包括“数据库操作”、“关联分析”、“预警操作”和“评估操作”4 大模块。下面分别对各大模块进行介绍。

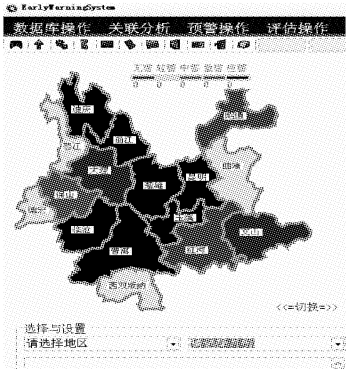


图 1　预警系统主界面

4.1.2 数据库操作模块

数据库操作模块包括“分割数据包”、“导入数据至内存”、“设置预测条件及挖掘”3 个操作。其中，“分割数据包”是将后台的源数据文件按地区分割成子块，然后将各个子块存放至相应地区所对应的后台数据库表中；“导入数据至内存”是将后台数据读到计算机内存，以供随后的挖掘操作使用；“设置预测条件及挖掘”是对已经在内存中的源数据进行挖掘操作。操作成功后将显示如图 2 所示的结果。

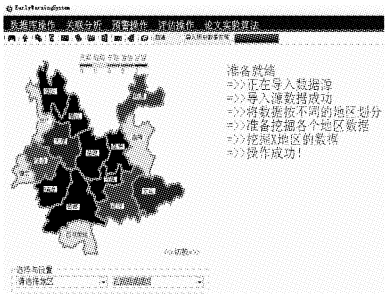


图 2 准备就绪提示界面

4.1.3 关联分析模块

关联分析模块用于计算指标间的相似度，如图 3 所示，图中显示的是指标“盗窃案件立案数”与其它指标的相似度，其中，与其关联最大的是指标“抢劫抢夺案件立案数”，关联程度高达 0.912。这里，关联程度最大为 1，最小为－1，负值表示负相关。

下面显示与指标“盗窃案件立案数”相似的指标及相似度值：

抢劫抢夺案件立案数 Sim= 0.912

刑事案件立案数 Sim= 0.807

治安案件发现受理数 Sim= 0.692

交通事故发生数 Sim= 0.645

刑事案件破案数 Sim= 0.446

每县的民警数 Sim= 0.144

图 3 指标间的关联分析

4.1.4 预警操作模块

预警操作模块包括“警戒级别阈值设置”、“显示预测结果及警戒水平”以及“显示预测结果的历史依据”。其中，“警戒级别阈值设置”由用户根据应用需要进行设置，比如某个指标的变化率大于某阈值，又或者某指标的预测数据值大于某个预设阈值，则将该指标的当前警戒状态设定在相应的级别。“显示预测结果及警戒水平”的结果如图 4 所示，例如，X 地区未来第一个月指标“盗窃案件立案数”的预测值为 756 件，相应的警戒水平为巨警。

地点	时间	盗窃案件立案数	交通事故发生数	每县的民警数	命案发生数
区地区	未来第1个月（数据）	756	458	851	3
	警戒水平=>>	巨警	巨警	无警	无警
X地区	未来第2个月（数据）	945	457	851	3
	警戒水平=>>	巨警	巨警	无警	无警
X地区	未来第3个月（数据）	700	449	852	3
	警戒水平=>>	巨警	巨警	无警	中警

图 4 预测结果及警戒水平示例

“显示预测结果的历史依据”的结果如图 5 所示。从图中可以看出，指标“盗窃案件立案数”在未来第 1 个月的警戒级别为巨警所依据的历史数据，其中影响最大的是 2011 年 10 月份该地区同类案件发生件数为 1444 件，占总影响因素的11.79%。

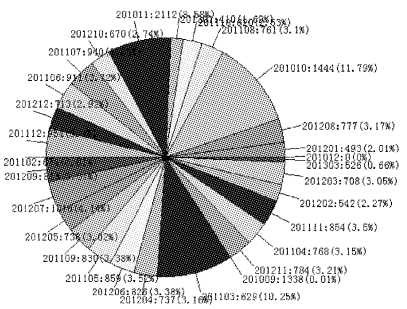


图 5 预测结果的历史依据

4.1.5 评估操作模块

“评估操作”用来测试滑动窗口模式匹配加权预测算法在各个不同大小训练集中进行预测的准确率，并检验各个指标的警戒阈值设置的合理性。如图 6 所示，设置训练集的初始大小占整个数据包的 80%，预测未来 4 个月的相关指标数据，随后训练集每次以 4 个月的数据量增加，以此类推。

图 6 评估操作参数设置

在此，对两算法的准确率评估是取所有训练集的相应准确率的平均值，如图 7 所示。

显示评估的平均准确率

指标	未来第 1 个月(平均准确率)	未来第 2 个月(平均准确率)	未来第 3 个月(平均准确率)
准确率计算量标准—基于量化法			
X地区 的平均准确率：	（准确率计算量标...	基于量化法）	
盗窃案件立案数：	0.925	0.883	0.622
交通事故发生数：	0.857	0.825	0.863
每县的民警数：	0.998	0.998	0.993
命案发生数：	0.699	0.242	0.483
抢劫抢夺案件立案数：	0.804	0.907	0.585
吸毒人数：	0.996	0.997	0.997
刑事案件立案数：	0.92	0.927	0.736
刑事案件破案数：	0.701	0.552	0.307
刑事案件致人死亡：	0.731	0.359	0.595
治安案件发现受理数：	0.563	0.654	0.852

图 7 各指标预测结果的准确率

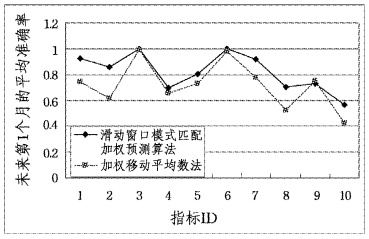
4.2 实验分析

实验分析所用的机器配置为：Intel(R) Celeron(R) CPU 1017U@1.60GHz 1.60GHz 处理器，4GB 内存，Win 7 操作系统。所用数据来自某城市在过去 6 年的社会治安状况指标数据记录。为进行有效的评估和分析，从源数据中随机抽取 10 个指标，训练集大小分别取 58 个月、62 个月、66 个月和 70 个月的历史数据，然后分别对这 4 个训练集的 10 个指标进行预测评估和分析。

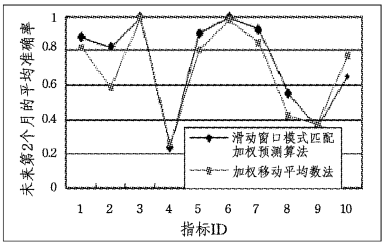
根据对几个数值预测算法的实验分析和比较，选取了预测效果较好的加权移动平均数法与本文所提出的预测算法进行了对比。

4.2.1 算法 1 的准确率

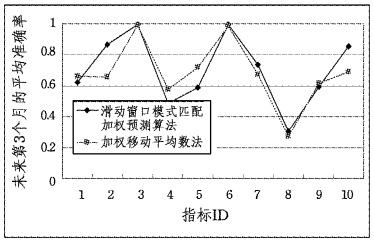
如图 8(a)所示，在未来第 1 个月中，针对 10 个指标进行预测，算法 1 的准确率总体水平明显高于加权移动平均数法。如图 8(b)所示，在未来第 2 个月中，有 5 个指标的滑动窗口模式匹配加权预测算法的准确率要明显高于加权移动平均数法，有 4 个指标两算法的准确率水平接近，而有 1 个指标用加权移动平均数法的准确率更高。在未来第 3 个月和第 4 个月中(图 8(c)和图 8(d))，算法 1 的总体准确率水平仍然略高于传统的加权移动平均数法，但优势并不明显。因此，与传统的加权移动平均数法相比，滑动窗口模式匹配加权预测算法在对近期的预测中表现更佳，特别是未来的前两个月，这对社会治安管理部门来说是一个尤为宝贵的信息。



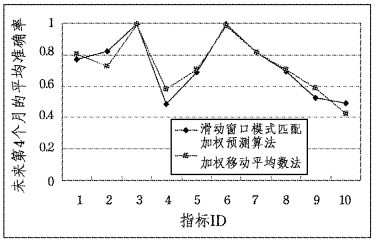
(a)未来第1个月的平均准确率



(b)未来第2个月的平均准确率



(c)未来第3个月的平均准确率



(d)未来第4个月的平均准确率

图8　未来不同时期不同指标的准确率评估

图9是未来4个月10个指标的平均准确率。从图中可以看出,在未来的前3个月中,滑动窗口模式匹配加权预测算法的准确率要高于传统的加权移动平均数法,特别是在未来的第1个月,前者的准确率比后者高出约10个百分点,最高达到81.9%,最低也达70.3%。

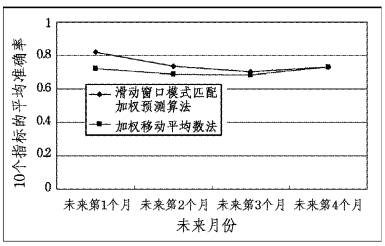


图9　未来4个月10个指标的准确率评估

4.2.2　加强算法的效果

本节对加强算法进行实验评估。由于该算法只针对个别预测准确率不佳的指标进行修正,因此实验也只针对这些特殊指标进行。

在训练集与预测集之比为9∶1的情况下,先用算法1考查对预测某地区未来3个月指标数据的准确率,结果如图10所示,其中指标"治安案件发现受理数"的准确率最差,3个月的预测准确率均为0。

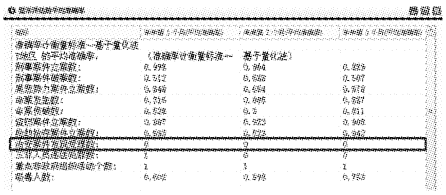


图10　算法1的预测准确率

为了对指标"治安案件发现受理数"的预测结果进行修正,提高其准确率,先查看与该指标相关性最大的其它指标,可发现指标"刑事案件破案数"、"命案侦破数"和"黑恶势力案件立案数"与该指标相关性最大。同时,从图10可知,这3个指标均具有相对较高的准确率。于是,执行加强算法得到指标"治安案件发现受理数"的修正结果,如图11所示。

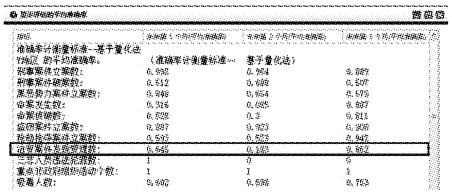


图11　指标"治安案件发现受理数"的修正准确率

4.2.3　时间效率分析

实验数据如表2所列。预测未来4个月的指标数据,比较算法的时间效率。如图12所示,滑动窗口模式匹配加权预测算法比传统的加权移动平均数法所消耗的时间要多,但二者的最大时间消耗都在两分多钟之内,由于前者具有更高的准确率,特别对近期的预测更是如此,因此滑动窗口模式匹配加权预测算法的综合性能更佳。

表2　本节实验数据

每行记录包含的待预测指标数(个)	34
数据集dataSet1的记录数(行)	500
数据集dataSet2的记录数(行)	1000
数据集dataSet3的记录数(行)	1500
数据集dataSet4的记录数(行)	2000

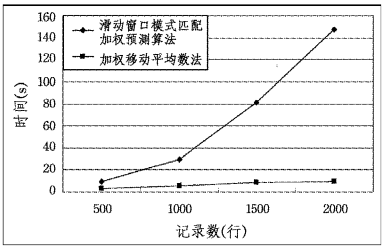


图12　算法的时间消耗

结束语　鉴于社会治安状况预测算法研究的价值和重要性,在尝试各种经典的数值预测方法无果的情况下,本文提出了一种简单、新颖的算法——滑动窗口模式匹配加权预测算法,用以预测未来某一时期的社会治安状况指标数据。大量的实验结果和实际部署表明,新算法具有较高的准确率,尤其是预测未来一两个月的情况,同时也具有较好的稳定性。在未来的工作中,将针对原历史数据中存在大量缺失数据的情况进行研究,以进一步提高算法的容错能力及准确率。

参 考 文 献

[1] Tan Pang-ning, Michael S, Vipin K. 数据挖掘导论[M]. 范明, 范宏建, 等译. 北京: 人民邮电出版社, 2006

[2] 王丽珍, 周丽华, 陈红梅, 等. 数据仓库与数据挖掘原理及应用(第二版)[M]. 北京: 科学出版社, 2009

[3] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology [J]. IEEE Trans. System, Man and Cybernetics, 1991, 21(3): 660-674

[4] 陈红梅, 王丽珍, 刘惟一, 等. 基于可达概率区间的不确定决策树[J]. 计算机科学与探索, 2012, 6(8): 726-740

[5] Ramoni M, Sebastiani P. Robust bayes classifiers [J]. Artificial Intelligence, 2001, 125(1/2): 209-226

[6] Schlkopf B, Smola A J. Learning with kernels: support vector machines, regularization, optimization, and beyond [M]. MIT Press, 2001

[7] 夏国恩, 金炜东. 基于支持向量机的客户流失预测模型[J]. 系统工程理论与实践, 2008, 28(1): 71-77

[8] Shih P C, Liu C J. Face detection using discriminating feature analysis and support vector machine in video[J]. Pattern Recognition, 2006, 39(2): 260-276

[9] Kim K, Jung K, Park S, et al. Support vector machines for texture classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(11): 1542-1550

[10] Shah C A, Watanachaturaporn P, Varshney P K, et al. Some recent results on hyperspectral image classification[C]// Proc. of 2003 IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data. Washington, DC, 2003: 346-353

[11] Neslin S A, Gupta S, Kamakura W, et al. Detection defection: measuring and understanding the predictive accuracy of customer churn models [J]. Journal of Marketing Research, 2006, 43(2): 204-211

[12] Kim E, Kim W, Lee Y. Combination of multiple classifiers for the customer's purchase behavior prediction [J]. Decision Support Systems, 2002, 34: 167-175

[13] 李元城, 方廷健, 郑国祥. 短期电力负荷预测的小波支持向量机方法研究[J]. 中国科学技术大学学报, 2003(12): 726-732

[14] Chang B R. Forecasting short-term stock price indexes-an integrated predictor vs. neural network predictor[C]// Proc. of 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. 2002: 817-820

[15] 温冬琴, 王建东. 基于奇异谱分析的机场噪声时间序列预测模型[J]. 计算机科学, 2014, 41(1): 267-270

[16] 赵泽亚, 贾岩涛, 王元卓, 等. 基于动态异构信息网络的时序关系预测[J]. 计算机研究与发展. 2015, 52(8): 1735-1741

[17] 王燕. 应用时间序列分析[M]. 北京: 中国人民大学出版社, 2008

[18] 孙丽梅, 李晶皎, 孙焕良. 基于动态 k 近邻的 SlopeOne 协同过滤推荐算法[J]. 计算机科学与探索, 2011, 5(9): 857-864

(上接第 577 页)

元数据的数据量比较小, 从而使得预取错误的代价比较低。当然当错误率提高时, 可能会超过预取带来的性能提升, 这时候系统考虑采用一个阈值来控制文件相关度的计算, 在实验中通过观察阈值对性能的影响来修改阈值的取值。

4.2 数据布局优化

文件相关性也被证明在提高文件数据布局策略的效果上有非常大的作用[7]。在分析文件相关性优化文件数据布局的过程中, 有若干问题需要考虑, 最重要的一个问题就是需要确定哪些文件应当被集成到一个文件组内。可以利用前面提到的相关性列表来解决该问题, 不过因为归档系统中文件不具备修改的特性, 因此数据布局相对来说显得较为简单, KingCloud 归档系统只考虑了只读相关文件来存放到同一个文件组内。这样, 不论是该组内的任何文件被访问, 整个访问组的文件数据都会被捆绑地读取到系统缓存, 从而为后续的访问提升 I/O 性能。

4.3 安全和可靠性策略

文件语义同样可以用来提升存储系统的安全和可靠性感知。例如 KingCloud 归档系统在安全删除和恶意访问拒绝策略中, 给定用户使用一个基于规则的文件和目录访问规则, 该规则被自动地应用到与这些文件具有很强相关性的文件和目录上, 也就达到不需要对后续归档文件再进行繁琐的角色分配或规则制定, 而是通过文档的相关度就能够直接达到智能安全访问策略的效果。

此外, KingCloud 归档系统在文件副本和对应的一致性管理中也使用了语义信息, 将相关文件作为一个集合放入一个逻辑副本组, 该逻辑副本组上的备份和恢复任务能够以原子操作的形式完成, 以保证同一副本组内文件的一致性。

结束语 本文描述了 KingCloud 智能对象归档系统的整体结构, 介绍了基于访问行为的语义采集方法, 随后对各种类型的非结构化文档的基于内容的元数据提取方法进行了详细分析。除此之外, 对于应用系统富含文档元数据描述的情况, 本系统也提供 ETL 工具对该部分数据进行抽取转换, 在这些工作的基础上, 分析了智能化的元数据获取技术对归档系统的元数据服务、对象数据布局优化、搜索效果优化等的性能有大幅的提升。

下一步的主要工作在于结合实际应用背景, 在应用系统数据 ETL、基本文档的元数据采集、文件与元数据关联方面得到具体的实践支撑, 继续丰富文件语义分析方法。

参 考 文 献

[1] Beaver D, Kumar S, Li H. Finding a needle in Haystack: Facebook's photo storage 2010[C]// Proc of the 10th USENIX Symp on Operating Systems Design and Implementation. 2010: 30-35

[2] 曹强, 黄建忠, 万继光, 等. 海量网络存储系统原理与设计[M]. 华中科技大学出版社, 2010

[3] Zeng L, Zhou K, Shi Z, et al. HUSt: a heterogeneous unified storage system for GIS grid[C]// Proceedings of SC '06: Proceedings of the 2006 ACM/IEEE Conference on Super Computing. New York, NY, USA: ACM, 2006: 325

[4] Semantic file systems [OL]. http://www.objs.com/survey/OFSExt.htm

[5] Bhagwat D, Polyzotis N. Searching a file system using inferred semantic links[C]// Proceedings of HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia. New York, NY, USA: ACM, 2005: 85-87

[6] Kroeger T M, Long D D E. The Case for Efficient File Access Pattern Modeling[C]// Proceedings of HOTOS '99. Washington, DC, USA: IEEE Computer Society, 1999: 14

[7] Agrawal N, Bolosky W J, Douceur J R, et al. A five-year study of file-system metadata[C]// Useninx Conference on File and Storage Technology. 2007