基于QPSO-LSSVM的数据库相似重复记录检测算法

梁 雪 任剑锋 景 丽
(河南财经政法大学计算机与信息工程学院 郑州 450002)

摘 要 针对大规模数据库的相似重复记录的检测问题,提出了一种量子群优化算法(QPSO)与最小二乘支持向量机(LSSVM)相结合的相似重复记录检测方法(QPSO-LSSVM)。首先计算记录字段的相似度值;然后利用QPSO对LSSVM参数进行优化,构建相似重复记录检测模型;最后通过具体数据集进行仿真测试实验。仿真结果表明,QPSO-LSSVM不仅提高了重复记录检测准确率,而且提高了检测效率,是一种有效的相似重复记录检测算法。

关键词 量子粒子群优化算法,最小二乘支持向量机,相似重复记录,检测
中图法分类号 TP393 **文献标识码** A

Approximate Duplicate Record Detection Algorithm Based on PSO and LSSVM

LIANG Xue REN Jian-feng JING Li
(School of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou 450002, China)

Abstract Approximately duplicate record detection algorithm was proposed based on quantum swarm algorithm (QPSO) and least squares support vector machine (LSSVM) to solve the large-scale database approximation duplicate record detection problem. Firstly, the record field similarity values are calculated, and then the LSSVM parameters are optimized, by QPSO to construction the approximately duplicate records detection model, finally simulation experiments are carried out on the data set. The simulation results show that QPSO-LSSVM not only improves the accuracy of the duplicate record detection but also improves the detection efficiency, and it is an effective approximate duplicate record, Detection algorithm.

Keywords Quantum particle swarm optimization, Least square support vector machines, Approximately duplicate record, Detection

1 引言

随着信息技术的发展,数据库集成在各个领域得到广泛应用。然而在数据库集成过程中,由于各种原因,如数据输入错误、不同表示方式等,导致数据库中包括一些噪声数据,因此必须对这些数据进行清洗处理,以提高数据质量,其中相似重复记录的检测是数据清洗中的一个关键步骤,故相似重复记录检测和清除成为数据库管理中的一个重要研究课题[1]。

针对相似重复记录检测问题,国内外学者对其进行了大量的研究,检测算法主要有:基于q-gram算法、距离函数模型、"排序+合并"的方法和字符串度量等方法[2-4],这些传统方法对于海量数据库中的相似重复记录进行检测时,空间复杂度和时间复度都比较大,最大的缺陷是它们假设字段和记录之间相似度是一种线性相关,而实际上,字段和记录之间相似度是一种复杂非线性关系,因此处理效率低,且难以获得较高的处理精度。近些年,随着机器学习算法不断成熟,出现了基于支持向量机、神经网络等的非线性数据库相似重复记录检测算法,从而有效提高了检测精度和效率[5,6]。但是它们都存在各自的不足,如对于大样本数据,支持向量机训练时间长,且计算复杂度高;神经网络学习速度快,但是结构复杂,参数难以确定,容易出现过拟合现象[7]。最小二乘支持向量机(Least Square Support Vector Machines, LSSVM)是一种改进的支持向量机,用等式约束代替传统支持向量机不等式约束,并且将二次规划方法求解过程变为解一组等式方程来加快训练速度,使其更加适合于海量数据库相似重复记录检测[8]。

为了更进一步提高数据库相似重复记录检测精度,充分利用粒子群算法全局寻优能力和LSSVM快速非线性映射能力,提出一种量子粒子群算法(Quantum Particle Swarm Optimization, QPSO)与LSSVM相融合的相似重复记录检测算法(QPSO-LSSVM)。仿真实验结果表明,QPSO-LSSVM不仅提高了相似重复记录检测精度,而且提高了检测效率,能够很好地解决海量数据库中的相似重复记录检测难题。

2 QPSO-LSSVM算法

2.1 LSSVM算法

支持向量机的复杂度与输入空间的维数无关,而依赖于样本数据的个数,因此样本数目越大,求解相应的二次规划问

到稿日期:2012-01-09 返修日期:2012-05-07 本文受河南省科学技术厅科技攻关科学项目(112102210199),河南省科学技术厅基础与前言研究项目(112300410201)资助。

梁 雪(1982-),女,讲师,主要研究方向为计算机网络;任剑锋(1979-),男,硕士,讲师,主要研究方向为软件工程;景 丽(1971-),女,博士,副教授,主要研究方向为人工智能。

题越复杂，计算速度越慢，从而限制了支持向量机的应用范围[9]。

Suykens 等在标准支持向量机的基础上提出了最小二乘支持向量机(LSSVM)，将标准支持向量机型中的损失函数设定成误差平方和，把不等式约束改成等式约束，减少待定参数，又将求解二次规划的问题转化成线性 KKT(Karush Kuhn Kucker)方程组的求解，降低了求解的复杂性，加快了求解速度[10]。

对于训练样本集$\{x_i, y_i\}$，$i=1,2,\cdots,n$，通过非线性映射函数 $\Phi(\cdot)$将样本映射到高维特征空间，从而获得最优线性回归函数：

$$f(x)=w^{\mathrm{T}}\varphi(x)+b \tag{1}$$

式中，ω 为特征空间的权值向量，b 为偏置量。

根据结构风险最小化原则，式(1)问题求解的 LSSVM 回归模型为：

$$\begin{aligned}&\min \ \| w \|^2+\frac{1}{2}\gamma\sum_{i=1}^{n}\xi_i^2\\&\text{s.t. } y_i-w^{\mathrm{T}}\varphi(x)+b=e_i,(i=1,2,\cdots,n)\end{aligned} \tag{2}$$

式中，γ 为惩罚参数；e_i 为实际值与回归函数间的误差。

通过引入拉格朗日乘子(Lagrange multiplier)将上述约束优化问题转变为无约束对偶空间优化问题，即：

$$L(w,b,\xi,\alpha)=\min \| w \|^2+\frac{1}{2}\gamma\sum_{i=1}^{n}\xi_i^2+\sum_{i=1}^{n}\alpha_i(w^{\mathrm{T}}\varphi(x)-b+e_i-y_i) \tag{3}$$

式中，α_i 为拉格朗日乘子。

根据 Mercer 条件，核函数定义如下：

$$K(x_i,x_j)=\varphi(x_i)^{\mathrm{T}}\varphi(x_j) \tag{4}$$

本文选择径向基核函数作为 LSSVM 核函数，径向基核函数为：

$$k(x_i,x_j)=\exp(-\| x_i-x_j \|^2/2\sigma^2) \tag{5}$$

最后 LSSVM 回归模型为：

$$f(x)=\sum_{i=1}^{N}\alpha_i \exp(-\| x_i-x_j \|^2/2\sigma^2)+b \tag{6}$$

式中，σ 表示径向基核函数宽度。

2.2 QPSO 对 LSSVM 参数优化

大量研究和实践表明，LSSVM 的预测性能与其参数关系十分密切，要获得高性能的 LSSVM 模型首先必须获得最适合的参数。由于当前 LSSVM 参数优化方法均存在各自不足，建模效果具有一定的局限性。量子粒子群优化(QPSO)算法对粒子群优化(PSO)算法进行了改进，具有简单易行、全局搜索能力强的优点，采用 QPSO 对 LSSVM 参数 γ 和 σ 进行优化可以提高 LSSVM 的预测性能。

2.2.1 QPSO 算法

在 QPSO 算法中，在 D 维搜索空间中共有 m 个粒子，第 i 个粒子的位置为：$x_i=(x_{i1},x_{i2},\cdots,x_{id})$，其搜索到的最优位置为：$P_{Besti}=(p_{bi1},p_{bi2},\cdots,p_{bid})$，粒子群的最优搜索位置为：$G_{best}=(bg_1,bg_2,\cdots,bg_d)$，粒子在第 d 维的值 pb_d 为：

$$pb_d=\frac{\phi_1\times p_{bid}+\phi_2\times g_{bg}}{\phi_1+\phi_2} \tag{7}$$

式中，ϕ_1、ϕ_2 为随机数。

那么，粒子的进化方程为[11]：

$$x(t+1)=pb\pm\beta\times|m_{best}-x(t)|\ln(1/\mu) \tag{8}$$

式中，m_{best} 表示粒子群最优的中值，β 表示调节算法的收敛速度。

当进行到第 t 次时，β 的值为：

$$\beta=b+c\times(T_{\max}-t)/T_{\max} \tag{9}$$

2.2.2 QPSO 对 LSSVM 参数优化过程

(1)根据问题规模确定粒子群规模 m，对粒子进行初始化处理，粒子的两维向量代表参数 γ 和 σ，以确定最大迭代次数和其它参数。

(2)采用 LSSVM 对训练样本进行训练建模，并计算粒子适应度值，适应度函数定义为：

$$f(i)=\frac{1}{E(i)} \tag{10}$$

$$E(i)=\sum_{p}\sum_{k}(V_k-T_k)^2 \tag{11}$$

式中，V_k 表示实际输出，T_k 表示期望输出。

(3)计算粒子的个体和全局极值，建立公共信息库。

(4)通过公共信息库对粒子群的历史最优个体和全局最优极值进行更新。

(5)对每一个粒子的位置采用式(8)、式(9)进行更新。

(6)对结束条件进行判断，如果满足结束条件，则转到步骤(7)，否则，增加迭代次数，返回到步骤(2)，重复迭代。

(7)对最优粒子位置进行反编码得到 LSSVM 的最优参数 γ 和 σ。

3 QPSO-LSSVM 的相似重复记录检测算法

3.1 字段相似度计算

在一个数据库中，语法上相同或相似的不同记录可能代表现实世界中的实体。由于数据库中的字段大多数都是字符串类型，而且重复记录之间的差异也是由字符串数据引起的，字段相似度计算是相似重复记录检测的基础，本文采用 Jaro 算法与 TF-IDF(Term Frequency-Inverse Document Frequency)算法相结合的字段相似度量计算方法。

3.1.1 Jaro 算法

Jaro 算法是 2 个字符串相似度的一种计算方法，其基本思想是根据 2 个字符串之间共有字符的数量和顺序对计算两者相似度，以正确识别拼写错误。对于待对比 2 个字符串 s_1，s_2，Jaro-Winkler distance 算法的相似度量函数为：

$$Jaro(s_1,s_2)=\frac{1}{3}\left(\frac{m}{|s_1|}+\frac{m}{|s_2|}+\frac{m-t}{m}\right) \tag{12}$$

式中，m 表示匹配的字符数，t 表示待换位的字符数。

匹配窗口(match window)计算公式为：

$$MW=\left(\frac{\max|s_1|,|s_2|}{2}\right)-1 \tag{13}$$

对于两个字符，如果 $Jaro(s_1,s_2)<MW$，那么认为两者匹配(相似)；如果两者匹配(相似)，但是字符位置不一样，那么换位的字符数(t)为不同顺序的匹配字符数目的一半。

3.1.2 TF-IDF 算法

设一个关系表 R，D 表示 R 中某一字段中所有的单词集合，s_1，s_2 表示 2 个字符串型的字段值，w 表示 s_1 中一个单词，那么根据 TF-IDF 算法可以得到 w 的权值：

$$V'(w,s_1)=\log(tf_w+1)\times\log(idf_w) \tag{14}$$

式中，tf_w 表示单词 w 在字符串 s_1 出现的次数，idf_w 表示单词 w 在单词集合 D 出现的频率。

则有：

$$\text{TF-IDF}(s_1,s_2)=\sum_{w\in s_1\cap s_2}V(w,s_1)V(w,s_2) \tag{15}$$

式中，$V(s_1)$表示字符串 s_1 中全部单词的权值向量。

字符串 s_1 和 s_2 的相似度量函数为：

$$\text{TF-IDF}(s_1,s_2)=\sum_{w\in s_1\cap s_2}V(w,s_1)V(w,s_2) \tag{16}$$

3.1.3 字段最终相似度量函数

将 Jaro 算法和 TF-IDF 算法的相似度量函数结合起来，

对于 $w\in s_1$，$v\in s_2$，则(w,v)表示集合 $close(\theta,s_1,s_2)$，从而得到最终字段相似度量函数为：

$$\mathrm{sim}(s,t)=\sum_{(w,v)\in close(\theta,s_1,s_2)} V(w,s_1)\cdot V(v,s_1)\cdot Jaro(w,v) \tag{17}$$

3.2 QPSO-LSSVM 相似重复记录检测

基于 QPSO-LSSVM 算法的数据库相似重复记录检测算法的基本思想为：首先通过 Jaro 算法与 TF-IDF 算法对记录字段的相似度进行计算，然后将字段相似度特征向量输入到 LSSVM 中进行学习，并计算整个记录对相似度，其分为两个阶段：训练阶段和检测阶段，具体如下：

(1)训练阶段。首先从数据集中选择若干个记录对组成训练样本集，然后采用 Jaro 和 TF-IDF 对字段相似度进行计算，得到记录对相对应字段的相似度值，同时对记录对相似值进行标记，最后将提取的相似度特征向量输入到 LSSVM 进行训练，训练过程采用 QPSO 对 LSSVM 的参数进行优化，从而得到数据库相似重复记录检测模型。

(2)检测阶段。对于测试样本集，采用训练样本相同的方法提取记录的字段特征向量，然后采用建立的相似重复记录检测模型对记录相似度值进行计算，选择一个阈值 δ，根据相似度值与 δ 的比较结果，来判断其是否重复记录。具体检测流程如图 1 所示。

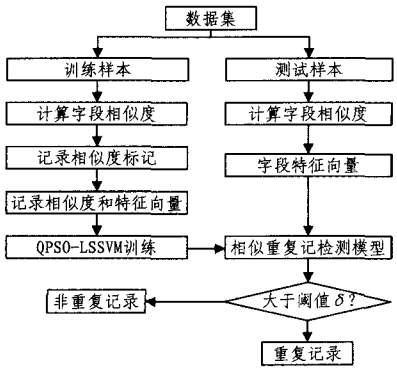

图 1 基于 QPSO-LSSVM 相似重复记录检测流程

4 仿真实验

4.1 仿真数据集

为了对本文数据库相似重复记录检测算法的性能进行检验，采用云南师范大学教务系统的学生选课数据库作为仿真数据集，随机选择 1000、5000、10000、30000 条记录，它们中的数据具体分布见表 1，训练样本和测试样本按 1:4 进行划分。

表 1 数据库中的记录分布情况

	总记录数	重复记录数
数据集 1	1000	150
数据集 2	5000	800
数据集 3	10000	1200
数据集 4	30000	2000

4.2 对比模型及性能评价标准

为了使检测结果具有说服力和可比性，采用标准支持向量(SVM)作为模型，模型性能评价指标为召回率(recall)和准确率(precision)以及训练时间。召回率和准确率定义如下：

$$Recall=\frac{TP}{RP}\times 100\% \tag{18}$$

$$precision=\frac{TP}{DP}\times 100\% \tag{19}$$

式中，TP 表示正确检测出的重复记录数，RP 表示重复记录总数，DP 表示所有检测出的重复记录数。

4.3 模型参数寻优

QPSO 算法的参数设置为：粒子群规模为 10，进化代数为 100，LSSVM 参数 γ 和 σ 的范围分别为[1,1000]和[0.1,100]。采用 Jaro 和 TF-IDF 组合算法计算记录字段的相似度值，并对训练样本中的记录进行标记，然后将训练样本特征向量值和记录标记输入 LSSVM 进行学习，QPSO 对参数进行优化，最后得到 $\gamma=10,\sigma=0.5$；采用 $\gamma=10,\sigma=0.5$ 对训练样本重新进行学习，建立最优数据库相似重复记录检测模型，同时采用 SVM 对训练样本进行学习，建立相应的相似重复记录检测模型。

4.4 实验结果与分析

采用上述建立的相似重复记录检测模型对 4 个数据集的测试样本集进行检测，得到的检测结果如表 2 所列。

表 2 两种重复记录检测算法的实验结果对比

数据集	QPSO-LSSVM		SVM	
	召回率	准确率	召回率	准确率
数据集 1	94.85	95.05	93.63	90.45
数据集 2	93.92	95.74	92.73	91.69
数据集 3	95.38	95.27	92.71	92.27
数据集 4	98.21	93.49	91.34	91.23

对表 2 两种算法的数据库相似重复记录检测结果进行对比可知，本文算法具有更好的检测准确率和召回率，尤其对大规模数据集进行检测时结果优劣更加明显，说明本文算法的鲁棒性更强，整体性能更优，具有更好的检测效果。

两种算法的训练时间如图 2 所示。从图 2 可知，本文算法检测时间要少于对比算法 SVM，尤其对于大规模数据集，训练速度更快，对比结果表明，QPSO-LSSVM 有效地提高了数据库相似重复记录检测效率，十分适应于实时性要求比较高的网络数据库重复记录检测。

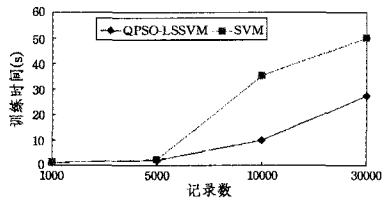

图 2 两种检测算法的训练时间对比

结束语 设计并实现了一种基于 QPSO-LSSVM 的数据库相似重复记录检测算法。仿真实验结果表明，QPSO-LSSVM 提高了重复记录检测正确率和检测速度，很好地解决了大规模数据库的相似重复记录检测问题。

参 考 文 献

[1] 郭志懋，周傲英. 数据质量和数据清洗研究综述[J]. 软件学报，2002,13(11):2076-2081

[2] Liang Jin, Chen Li, Mehrotra S. Efficient record linkage in large data sets[C]//Proc. of the 8th Int'l Conf on Database. Systems for Advanced Applications. Washington: IEEE Computer Society, 2003:137-148

(下转第 190 页)

为止。此时，新鲜度为 0 的信息簇可以直接作为深翻依据进行处理；而新鲜度为 1 的信息簇，则需要搜索是否有内外客户的及时需求，其最终将会被丢弃或者作为深翻依据。

5 实验结果与分析

本模型的验证实验在某高校内门户网站的“相关新闻”和 BBS 中的“校园轶事”等版块中进行，其中主观测试（通过对普通用户进行问卷调查）调查的普通用户人数为 271 人，回收有效调查问卷 229 份；调查的版主及管理员人数为 33 人，回收有效调查问卷 32 份。实验主要从用户的心理满足程度和信息的相关程度等内容对模型性能进行了测试，另外还对该模型对系统性能的“扰动”进行了测量，实验结果（与未用之前的测量结果进行对比）如图 3 所示。

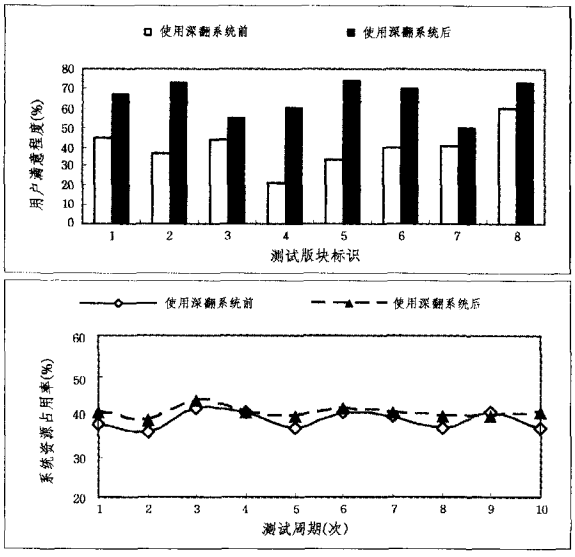

图 3 深翻系统仿真实验结果

图 3(上)中的测试结果显示使用深翻系统后，网站点击率明显提高，且用户对网站中的各版块用户满意度均有不同程度的提高，即深翻系统能够较好地从内外两个方向上获取用户的当前心理，并根据当前用户的心理需求从历史数据中“深翻”出相关度较高的内容。此外，版主与网站管理员的问卷调查也显示：深翻系统能够提供与用户心理匹配较好的信息列表，且自动化程度较高，能够节省大量的人力资源。

图 3(下)中显示了 10 个监测周期内系统的平均负载变化（网站流量、内存与 CPU 为主要监测对象），其中，测试环境采用了联想万全 R520 服务器和 Windows2003 操作系统。由图中可知，使用了深翻系统之后，网站系统的整体负载并未大幅度攀升，与使用前基本持平；通过进一步的分析，可究其原应为：由于深翻系统减轻了用户的信息检索工作量和人机交互通信量，使得网站系统占用的系统负载下降，部分抵销了网站内部“深翻”所产生的系统负载。

结束语 基于互联网用户心理挖掘的网站深翻模型在实际应用中取得了良好的效果，具有一定的推广价值。进一步的研究工作包括：多信息源辅助的深翻决策模型、海量和多批次的反向心理线索发掘模型等，并将其应用于公开股票信息挖掘等领域。

参 考 文 献

[1] 王辉，王晖昱. 观点挖掘综述[J]. 计算机应用研究，2009，26(1)：25-29

[2] 杨超，冯时，等. 基于情感词典扩展技术的网络舆情倾向性分析[J]. 小型微型计算机系统，2010，31(4)：691-695

[3] 张顺香，朱广丽，陆奎. 基于 Web 挖掘的主页多主题更新模型[J]. 计算机应用，2009，29(10)：2796-2799

[4] 李晓亚，赫枫龄，左万利. 基于网页分块技术主题爬行器的实现[J]. 吉林大学学报，2007，45(6)：959-965

[5] 阮光册. 基于兴趣度策略的启发式 Web 挖掘算法[J]. 计算机工程与应用，2009，45(35)：148-150

[6] 章剑锋，张奇，吴立德. 中文观点挖掘中的主观性关系抽取[J]. 中文信息学报，2008，22(2)：55-60

[7] 葛育祥，熊励. 整合文本挖掘的商务智能系统结构研究[J]. 计算机技术与发展，2009，19(4)：1-4

[8] 杨频，李涛，赵奎. 一种网络舆情的定量分析方法[J]. 计算机应用研究，2009，26(3)：1066-1069

[9] 周红芳，冯博琴，等. 基于语义模型的 Web 挖掘算法研究[J]. 哈尔滨工业大学学报，2009，41(11)：212-214

[10] 查凯莱巴蒂. Web 数据挖掘[M]. 北京：人民邮电出版社，2009：23-82

[11] Chou P-H，Li P-H. Integrating Web mining and neural network for personalized e-commerce automatic service[J]. Expert Systems with Applications，2010(37)：2898-2910

[12] Hung S-H，Lin C-H，et al. Web mining for event-based commonsense knowledge using lexico-syntactic pattern matching and semantic role labeling[J]. Expert Systems with Applications，2010(37)：341-347

（上接第 159 页）

[3] Elmagarmid A K，Panagiotis G，et al. Duplicate record detection：a survey[J]. IEEE Transactions on Knowledge and Data Engineering，2007，19(1)：1-16

[4] 张昌年. 一种基于 VSM 的检测相似重复记录的方法[J]. 微电子学与计算机，2008，25(8)：184-187

[5] 马翔. 粒子群优化 BP 神经网络用于重复记录检测[J]. 辽宁工程技术大学学报：自然科学版，2010，29(5)：959-963

[6] 朱恒民，王宁生. 一种改进的相似重复记录检测方法[J]. 控制与决策，2006，21(7)：805-813

[7] Elmagarmid K，Panagiotis G. Duplicate record detection：a survey [J]. IEEE Transaction on Knowledge and Data Engineering，2007，19(1)：1-16

[8] Minton S N，Nanjo C，Knoblock C A. A heterogeneous field matching method for record linkage[C]//Proceedings of the 5th International Conference on Data Mining. Washington：IEEE Computer Society，2005：314-321

[9] 巩知乐，张德贤，胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真，2009，26(7)：165-168

[10] 杨福刚. 基于人工免疫算法的最小二乘支持向量机参数优化算法[J]. 计算机应用研究，2010，27(5)：1702-1704

[11] 李旭芳，王士同. 基于 QPSO 训练支持向量机的网络入侵检测[J]. 计算机工程与设计，2008，29(1)：34-36

[12] 吴文欢，张少辉，李巍. 分阶段进化的粒子群优化算法[J]. 重庆理工大学学报：自然科学，2012，26(6)：67-70