文章编号：2095-2163(2022)09-0149-05 中图分类号：TP301.6 文献标志码：A

基于改进模糊 FP-Growth 的异常检测算法

杜嘉伟，余 粟
（上海工程技术大学 机械与汽车工程学院，上海 201620）

摘 要：本文提出一种基于改进模糊 FP-Growth 的异常检测算法-RFPG 算法（Random Frequency Pattern Growth），算法建立 2 层 FP-Tree。第一层基于 bagging 思想，随机采样生成集合并得到长频繁项集合；第二层将长频繁项集合作为输入，得到模式强关联规则集，再通过相似度计算进行异常检测分类。实验结果显示，本文提出算法的整体异常检测效率与质量良好。

关键词：异常检测；FP-Growth；bagging；关联规则

Anomaly detection algorithm based on improved fuzzy FP-Growth

DU Jiawei, YU Su
(School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 An anomaly detection algorithm based on improved fuzzy FP-Growth is proposed for the application of anomaly detection. The algorithm establishes two layers of FP-Tree. The first layer randomly samples based on bagging idea to generate a set and obtain a long frequent item set. The second layer takes the long frequent item set as the input to obtain a pattern association rule set, and then classifies the anomaly detection through similarity calculation. The experimental results show that the overall anomaly detection of the proposed algorithm is very effective. The measurement efficiency and quality are excellent.

【Key words】 anomaly detection; FP-Growth; bagging; association rules

0 引 言

随着大数据时代的来临，数据挖掘对于研究事务之间的关联规律模式有着重要的意义。目前，异常检测被广泛应用于工业控制系统、网络入侵检测、医疗异常检测等领域[1-4]。异常检测是基于当前数据模式检测出其中的异常点，即检测出不符合预期模式的数据。相比较于正常数据簇，异常值数据有着与数据集中正常数据并不相同的数据特征的数据模式。

近年来，国内外学者对于异常检测的研究已取得显著成果。其算法主要可以分类 4 类：基于模式预测的算法模型、基于机器学习的算法模型、基于统计的算法模型、基于贝叶斯网络的算法模型[5]。

文献[6]提出一种基于模糊孤立森林算法的多维数据异常值检测算法。该算法通过挑选一些有价值的属性建立异常检测模型，再通过隶属度判断得到评价结果，但选取的属性只能为类别型特征。文献[7]基于模糊理论，提出一种挖掘连续型数值数据的关联规则，但该算法基于正常模式的数据，无法处理黑盒数据。由此可见，上述方法大多都基于正常模式下的数据挖掘，在建立模型时要求其数据全为正常模式数据，并且要包含绝大多数的正常模式行为。在异常检测中，将没有标签标识的元数据称为黑箱状态的数据。在面对数据黑箱场景下，降低算法的误报率和漏检率，以及提高算法模型的泛化能力就成为了当下研究的热点与难点。

综上所述，本文提出一种基于改进 FP-Growth 算法的异常检测算法-RFPG 算法（Random Frequency Pattern Growth）。对于传统数据挖掘方法中以阈值对数值型特征数据进行划分遇到的尖锐边界问题[8]，本文提出一种基于模糊集理论的异常检测方法；对于数据黑箱问题，算法中基于 baggging 思想对数据随机采样，生成 FP-Tree 集合，挖掘出模式长频繁项集合，并通过将长频繁项集合作为输入生成新 FP-Tree，挖掘模式强关联规则集，使算法对于黑箱数据有着更好的寻优与泛化能力。

1 基本概念

1.1 模糊集和隶属度

模糊集是用来描述具有模糊语义集合概念。如果存在论域 U，而且其中任何一个元素 x 与 $A(x) \in$

基金项目：国家科技部“十二五”支撑计划项目“电子产品精密装配自动化生产线研制与示范”（2015BAF10B00）。

作者简介：杜嘉伟（1995-），男，硕士研究生，主要研究方向：数据挖掘；余 粟（1962-），女，博士，教授，主要研究方向：网络安全。

通讯作者：余 粟 Email：suyu_sh@hotmail.com

收稿日期：2021-11-13

[0,1] 存在对应关系,则 A 称为 U 上的模糊集,$A(x)$ 称为 x 对 A 的隶属度。

1.2 频繁项集

采用向量和矩阵的概念,建立频繁项集(Frequent itemsets)的挖掘模型。令 $I = \{item_1, item_2, \ldots, item_n\}$ 为 n 个数据项的集合。其中,$item_j(1 \leqslant j \leqslant n)$ 称为项 $D = \{t_1, t_2, \ldots, t_m\}$,是 m 个事务组成的数据集,$t_i(1 \leqslant i \leqslant m)$ 表示一条事务,且 t_i 是 I 的子集。

项的集合称为项集(*Itemset*),包含 k 个项的项集称为 k 项集。当一个 k 项集计算所得的支持度满足设定的支持度阈值时,则定义该项集为频繁 k 项集。

支持度(*Support*)描述了多个项集在所有事务中同时出现的概率。事务数据集中的支持度用 $Sup(X)$ 表示,支持度阈值用 min_sup 表示。在挖掘频繁项集的过程中,min_sup 能够筛选并枚举出所有重要的项集。基于以上描述可推得如下定义:

定义 1 给定 min_sup,若 $Sup(X) > min\_sup$,则项目集 X 是频繁项目集。

定义 2 存在项目集 X_1、X_2,而且 $X_1 \subseteq X_2$,则 $Sup(X_1) \geqslant Sup(X_2)$。

1.3 FP-Growth 算法

FP-Growth 算法是一种频繁项集挖掘算法,该算法中使用一种 FP 树的数据结构作为存储结构。在面对大量输入的情况下,树存储结构可以应对复杂的数据存储问题。算法通过以频繁一项集的支持度,倒序读取事务,并将其映射到 FP 树中。考虑到选用的存储结构,当大量读取相同事务的情况下,已构造的重叠路径能大大减少存储时间,并减少了空间上的开销。

FP-Tree 的存储结构避免了如 Apriori 算法那样频繁扫描数据库的 I/O 瓶颈,减少了 I/O 损耗,有着更优的时间与空间复杂度。FP-Growth 算法的主要任务是找出数据集中的频繁项集,算法的主要实现步骤可依次表述为:数据采样;构建 FP-Tree;基于 FP-Tree 挖掘长频繁项集;生成模式关联规则集;模式匹配。FP-Growth 算法伪代码详述如下。

输入: 网路流量数据集 D,min_sup

输出: 输出异常检测分类结果

```
//对流量特征进行标准化
1: D = MinMaxScaler(D)
//对数值特征进行模糊化处理
2: new_D = Fuzzy-C-Means(get_num_col(D))
3: while n <= n_estimator
     // k 为采样百分比
4:   sample_data = get_samples(k, new_D)
     //构建 fp-tree
5:   fptree_set = get_fptree(sample_data)
     //生成长频繁项集
6: freqset = freqset + get_freqset(fptree_set)
7: End while
   //生成关联规则集
8: association_rules_set = get_association_rules(freqset)
   //计算数据与关联规则集相似度
9: similarity_score = cal_similarity(x, association_rules_set)
   //相似度匹配
10: res = similarity_match(similarity_score)
11: return res
```

2 RFPG 算法

在传统 FP-growth 基础上,基于模糊理论的 RFPG 算法根据频繁 k 项集的隶属度,以降序方式生成条件模式基。每个条件模式基在经过剪枝后,获得候选 $k+1$ 项集,满足支持度要求的候选 $k+1$ 项集即为频繁 k 项集。于是提出如下定义:

定义 3 若条件模式基中存在 n 个满足 min_sup 的元素,则该条件模式集基中存在最长候选 $n+1$ 项集。

2.1 数据模糊化处理

基于模糊数学理论的综合评价方法,针对连续型数据特征进行数据模糊化处理。具体来说,模糊综合评判就是将数据模糊关系与模糊数学相关联,将难以确定边界的特征因素以隶属度的形式进行描述。

在挖掘包含大量连续数值型特征数据的关联规则时,由于传统 FP-Growth 算法只能处理二值化特征,无法处理连续数值型特征,因此引入数据模糊化处理改进 FP-Growth 算法,使算法的泛化能力更好,并能避免传统基于阈值划分时数据分类带来的尖锐边界问题。本文使用 Fuzzy-C-Means 算法对连续数值型数据进行模糊聚类,使用隶属度描述聚类程度。目标函数定义为:

$$\min J = \sum_{i=1}^{c} \sum_{j=1}^{n} w_{ij}^{m} (x_j - c_i)^2$$

$$\text{s.t.} \sum_{i=1}^{c} w_{ij} = 1, w_{ij} \in [0,1], j = 1, \ldots, N,$$

$$0 < \sum_{j=1}^{n} w_{ij} < n \tag{1}$$

通过公式(2)计算得到聚类中心点和目标函数值，并由公式(3) 重新计算权重矩阵 $\boldsymbol{W}$，当目标函数值小于误差阈值结束迭代。这里用到的公式为：

$$C_i = \frac{\sum_{i=1}^{N} w_{ij}^{m} X_i}{\sum_{i=1}^{N} w_{ij}^{m}} \tag{2}$$

$$w_{ij} = \frac{1}{\sum_{s=1}^{K} \left(\frac{\| X_i - C_j \|}{\| X_i - C_S \|} \right)^{\frac{2}{m-1}}} \tag{3}$$

将元数据集转化为模糊数据集，见表 1。表 1 中，d_{ij} 表示 FCM 聚类后结果。

表 1　模糊数据集 *D*

Tab. 1　Fuzzy dataset *D*

	$\omega_{11}\omega_{12}\ldots\omega_{1j}$	$\omega_{21}\omega_{22}\ldots\omega_{2j}$	…	$\omega_{k1}\omega_{k2}\ldots\omega_{kj}$
t_1	d_{11}	d_{12}	…	d_{1k}
t_2	d_{21}	d_{22}	…	d_{2k}
…	…	…	…	…
t_n	d_{n1}	d_{n2}	…	d_{nk}

模糊集的隶属度使用模糊集各属性的支持度计数所得，对此可表示为：

$$sup(\omega_{ij}) = \left(\sum_{m=1}^{n} \omega_{mij} \right) / n \tag{4}$$

计算得出各个项集的支持度后，对候选 k - 项集进行剪枝操作，得到频繁 k - 项集。其中，剪枝包含 2 个步骤，可做阐释分述如下：

(1)删除不满足 *min_sup* 阈值的候选项，得到频繁 k - 项集。

(2)删除包含非频繁 (k - 1) - 项集的候选项集，得到频繁 k - 项集。

2.2　异常检测

在进行异常检测之前，通常先挖掘出数据正常模式下的关联规则，但处于数据黑盒状态下的数据由于异常数据扰动，会使挖掘出的关联规则支持度与置信度会降低。因此，使用 RFPG 算法挖掘生成 FP-Tree 时，可以设置稍低的最小支持度与最小置信度阈值，以获得尽可能多的关联规则集；将各个 FP-Tree 集合挖掘出的关联规则集作为第二层输入，生成新的 FP-Tree，由此得到接近正常模式的关联规则集。

在数据中，由于异常数据簇与正常数据簇分布不同，RFPG 算法使用背离度这一概念来描述异常状态。通过规则集与单时间段数据的相似度，量化当前数据集与数据正常模式的背离度，以此建立异常检测系统的检测机制。算法中将模糊化处理后的单时间段数据与挖掘出的正常模式关联规则的交集作为单条数据的长频繁项集。对于关联规则相似度的计算，设存在关联规则 R_1 与某一时间段数据 D_1，这里 R_1 的支持度为 Sup_1，置信度为 $Conf_1$，则其关联规则的相似度可由式(5)计算求出：

$$S(R_1, D_1) = 1 - \max\left(\frac{|sup_1 - 1|}{sup_1}, \frac{|conf_1 - 1|}{conf_1} \right) \tag{5}$$

对于关联规则集相似度的计算，设存在关联规则集 S_1 与某一时间段数据 D_1。其中，规则集 S 包含若干关联规则 R_n，N_1 为关联规则集规则数量，N_2 为某一时间段数据元素与关联规则集元素交集个数。关联规则集的相似度的计算可写为式(6)：

$$S(S_1, D_1) = \frac{\left(\sum_{\substack{\forall R_1 \in S_1 \\ \forall R_2 \in S_2}} S(R_1, D_1) \right)^2}{N_1 N_2} \tag{6}$$

3　实验结果及分析

实验使用 NSL-KDD 网络流量数据，以评价 RFPG 算法异常检测的性能。实验环境为：64 位的 Win10 操作系统，Intel® CoreTM i7-9700kf CPU，主频 3.6 GHz、内存 32 GB；仿真编程语言为 Python3.8。

3.1　数据集介绍

NSL-KDD 数据集是改进 KDD-CUP99 的网络流量数据集，优化了数据集冗余和重复数据多的不足。数据为 9 个星期的网络连接数据，采集于模拟的某国空军局域网。

数据集有 41 个特征，包含网络连接特征和网络流量特征。标签列分为了 2 类：正常模式的标识类型(normal)和异常模式的标识类型(abnormal)。数据集共 125 972 条数据，样本分布见表 2。

表 2　NSL-KDD 数据集样本分布

Tab. 2　Samples distribution of NSL-KDD dataset

类别标签	类别	样本数	占比/%
0	正常数据	67 342	53.5
1	异常数据	58 630	46.5

3.2　评价指标

在异常检测中，误报率 (*FAR*) 与漏检率(*DR*)通常是衡量异常检测算法的指标[9]，而误报率与检测率是基于混淆矩阵的概念实现。混淆矩阵见表 3。对此拟做探讨阐述如下。

表 3 混淆矩阵

Tab. 3 Confusion matrix

	预测(异常)	预测(正常)
实际(异常)	*TP*	*FN*
实际(异常)	*TP*	*TN*

(1)误报率。是指正常而被误判为入侵样本数占正常样本数的百分比,数学定义公式如下:

$$FAR = \frac{FN}{FN + TN} \tag{7}$$

(2)检测率。是指异常而被误判为正常的样本数占异常样本数的百分比,数学定义公式如下:

$$DR = \frac{TN}{FN + TN} \tag{8}$$

3.3 实验过程

为了模拟真实网络攻击状况、并检验 RFPG 算法异常检测效果,实验分别准备了 3 组不同数据量的数据,分别为:5 000 条数据、10 000 条数据、15 000条数据。每组中使用 3 类数据,分别是:包含5%异常样本数据的数据集;包含 10%异常样本数据的数据集;包含 15%异常样本数据的数据集。

实验中,RFPG 算法使用前文求得的模糊聚类中心模糊化网络流量数据,按照先验知识将流量特征等数值特征 *num_col* 分为低、中、高三个模糊分区,并将连接特征等类别特征 *cat_col* 按照其类型数量进行编码,研究获得($a * num\_col + b * cat\_col$)个特征。其中,$a$ 为模糊分区数,b 为类别特征包含的类别数。在挖掘模式强关联规则时,设最小支持度 $min\_sup = 0.8$,最小置信度 $min\_conf = 0.8$。

数据量为 5 000、10 000、15 000 时 RFPG 算法的 *AUC* 评分结果曲线如图 1~3 所示。由图 1~3 可见,RFPG 算法整体异常检测效果良好,并且随着数据量的提高,包含不同比例的异常样本数据的分类 *AUC* 评分越稳定。由此证明:数据量越大,数据中包含的正常行为模式越完整,在挖掘关联规则时,实际为正常样本被预测为异常样本的概率降低,分类效果更好。由图 3 还可见到,RFPG 算法实验结果中,包含不同比例异常数据的分类结果评分相近,证明该算法抗噪声干扰能力强,对于黑盒数据的模型泛化能力也很强。

3.4 对比试验

实验使用基于 K-means 异常检测算法[8]、基于 KL 距离的异常检测算法[10]、Fuzzy-Apriori 算法[7]以及 RFPG 算法进行结果对比。

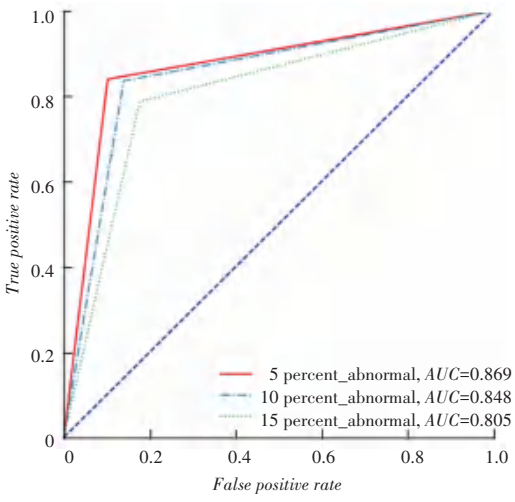

图 1 5 000 数据量 RFPG 算法 *AUC* 评分

Fig. 1 *AUC* score of RFPG algorithm when the amount of data is 5 000

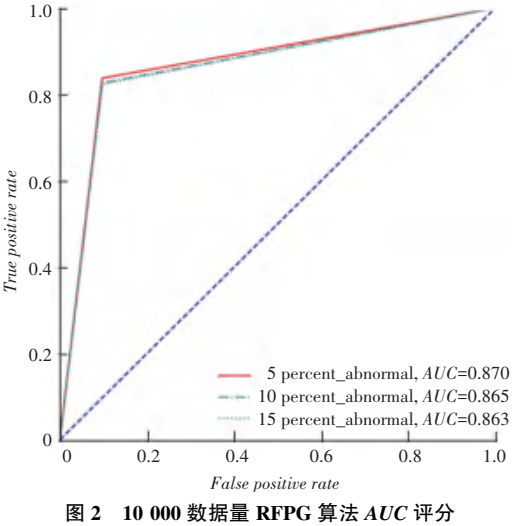

图 2 10 000 数据量 RFPG 算法 *AUC* 评分

Fig. 2 *AUC* score of RFPG algorithm when the amount of data is 10 000

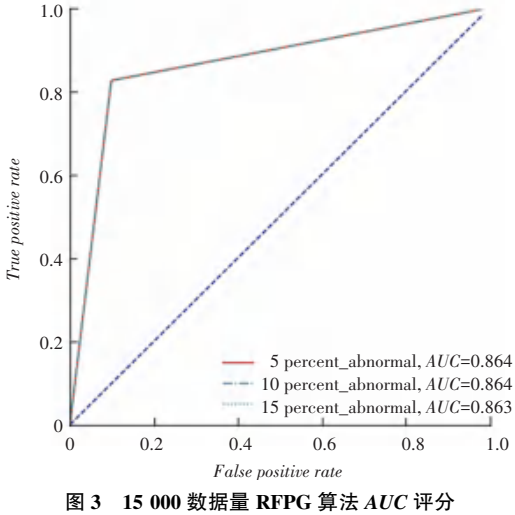

图 3 15 000 数据量 RFPG 算法 *AUC* 评分

Fig. 3 *AUC* score of RFPG algorithm when the amount of data is 15 000

对比实验选取同一包含 15 000 数据量以及 10% 异常样本的数据集，对 RFPG 算法的可用性进行验证，见表 4。结果表明，在实验数据相同的情况下，RFPG 异常检测算法相比其它算法，在误报率与检测率上都有着更好的检测表现。基于 K-means 的异常检测算法，对于数据特征变化偏移较小的异常数据容易产生漏检的情况；基于 KL 距离的异常检测算法，结合了 EWMA 预测模型刻画出了数据整体变化趋势，敏感性较高，有着较好的检测率，同时误报率也有所提升；Fuzzy-Apriori 算法基于模糊理论，根据数据分布进行隶属度函数的建立，对于有标签的数据有着较好的表现，但是对于黑箱数据检测率与误报率表现欠佳，并且时间复杂度较高。RFPG 算法基于 bagging 随机采样数据生成 FP-Tree 集合，得到长频繁项集合，并将其作为第二层的输入，生成近似正常模式的关联规则集，使算法面对黑盒数据依然有较为良好的表现，更符合异常检测的现实需求。同时，RFPG 算法基于 FP-Tree 结构，使算法相较于 Fuzzy-Apriori 算法有着较优时间复杂度。实验结果显示，较高的检测率与较低的误报率也验证了 RFPG 异常检测算法的可用性。

表 4　多算法对比试验

Tab. 4　Algorithms comparison experiment　%

异常检测算法	误报率	检测率
基于 K-means 异常检测算法	3.4	93.8
基于 KL 距离的异常检测算法	2.9	98.3
Fuzzy-Apriori 算法	5.5	91.4
RFPG 算法	1.1	98.9

4　结束语

本文提出的基于改进模糊 FP-Growth 异常检测算法 RFPG，经实验表明，在大数据体量下，其异常检测能力更优，并且由于 RFPG 基于树的存储结构更适用于大数据集体量的场景，减少了提取关联规则集的运行时间和内存消耗。但基于先验知识，在对数值型特征进行模糊化处理时，会增加数据集的维度。因此，需要对自适应隶属度函数的建立，以及在算法过程中动态削减数据集维度与优化搜索空间进入深入探讨，以达到进一步优化整体异常检测质量与效率的研究目的。

参考文献

[1] ZHOU Peng, XIONG Yunyu. Network state anomaly detection based on data mining[J]. Journal of Jilin University (science edition), 2017(5): 1269-1273.

[2] ZHENG Liming, ZOU Peng, JIA Yan. Research on the extraction and training methods of classifiers in network traffic anomaly detection[J]. Journal of Computer Science, 2012(4): 77-87, 185.

[3] CAI Ruichu, XIE Weihao, HAO Zhifeng. Population anomaly detection based on multi-scale time recurrent neural network[J]. Software Journal, 2015, 26(11): 140-152.

[4] QU Ping. Research on a new anomaly intrusion detection system based on data mining technology[J]. Application of electronic technology, 2010(8): 152-156.

[5] ZHANG Yang, MERATNIA N, HAVINGA P. Outlier detection techniques for wireless sensor networks: A survey[J]. IEEE Communications Surveys & Tutorials, 2010, 12(2): 159-170.

[6] 李倩，韩斌，汪旭祥. 基于模糊孤立森林算法的多维数据异常检测方法[J]. 计算机与数字工程，2020，48(04)：862-866.

[7] 熊平，朱天清，黄天戍. 模糊关联规则挖掘算法及其在异常检测中的应用[J]. 武汉大学学报(信息科学版)，2005，30(09)：841-845.

[8] 陈庄，罗告成. 一种改进的 K-means 算法在异常检测中的应用[J]. 重庆理工大学学报(自然科学)，2015，29(05)：66-70.

[9] 冯兴杰，焦文欢. 基于动态阈值的网络异常检测[J]. 计算机工程与设计，2012，33(06)：2182-2186.

[10] 蒋华，张红福，罗一迪，等. 基于 KL 距离的自适应阈值网络流量异常检测[J]. 计算机工程，2019，45(04)：108-113，118.

(上接第 148 页)

[2] 刘俸材，李爱迪，马泽忠. 结构光视觉系统误差分析与参数优化[J]. 计算机工程与设计，2013，34(02)：757-761.

[3] TÕLGYESSY M, DEKAN M, CHOVANEC L, et al. Evaluation of the azure kinect and its comparison to kinect V1 and kinect V2[J]. Sensors, 2021, 21(2): 413.

[4] TU Lifen, PENG Qi, LI Yang. Method of using realsense camera to estimate the depth map of any monocular camera[J]. Journal of Electrical and Computer Engineering, 2021, 2021(11): 1-9.

[5] 王胤. 应用于三维成像飞行时间法建模及其误差分析[D]. 湘潭：湘潭大学，2017.

[6] 张仲楠，霍炜，廉明，等. 基于 Yolov5 的快速双目立体视觉测距研究[J]. 青岛大学学报(工程技术版)，2021，36(02)：20-27.

[7] Hagadone Z. Leap motion controller[EB/OL]. [2013]. https://www.leapmotion.com/.

[8] ZHANG Z. A flexible new technique for camera calibration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330-1334.

[9] CAO Zhe, SIMON T, WEI S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 7291-7299.

[10] KREISS S, BERTONI L, ALAHI A. Pifpaf: Composite fields for human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 11977-11986.

[11] 郑军. 四元数插值算法实现游戏角色平滑旋转[J]. 阴山学刊(自然科学版)，2012，26(01)：14-15.

[12] 杨鹏. 多子群分层粒子群差分算法在机器人逆运动学中的应用[D]. 长沙：湖南大学，2015.

[13] SINGH G K, CLAASSENS J. An analytical solution for the inverse kinematics of a redundant 7DoF Manipulator with link offsets[C]// IEEE/RSJ International Conference on Intelligent Robots & Systems. Taipei: IEEE, 2010: 2976-2982.