基于疟疾传播动力学模型、多维空间GeoHash编码和交叠分区策略的并行化DBSCAN方法在犯罪团伙识别中的应用

## 背景技术

随着社会经济的快速发展，电信诈骗犯罪的发案率不断升高，相对于一般的刑事案件，电信诈骗犯罪一般为团伙作案，且可以借助方便的通信设备作案，对社会负面影响范围更广。如何遏制和打击电信诈骗犯罪团伙已经成为各地公安机关当前工作的重点任务。通过数据挖掘对相关人员进行聚类分析，可以有效地识别出犯罪团伙，该方法通常是从大量数据中挖掘出相关人员的特征信息，通过统计学技术识别出相关人员所处的犯罪团伙，从而帮助公安机关实现打击罪犯团伙的目的。但随着大数据时代的发展，人们的行为产生的数据特征和数据量快速增加，已经远远超出传统方法的数据处理能力。因此，传统的电信诈骗犯罪团伙识别方法并不能快速有效地在大数据环境下对犯罪团伙进行有效地甄别。为了更好的根据公安部门所掌握的数据来识别出电信诈骗犯罪团伙，创立一种支持多维度数据，高效，负载均衡的电信诈骗团伙甄别方法是很有必要的。

近年来出现了很多针对电信诈骗团伙识别的研究。例如，潘芳等人提出了一种基于模糊层次聚类算法的犯罪网络分析方法；李亮利用快速分裂算法对已经建立的嫌疑人社会网络进行分解，来缩小可能的犯罪团伙排查范围。张玉超提出空间轨迹向量化方法，最终利用K-means聚类算法来实现人员聚类的目的。蔡霖翔在K-means聚类算法基础上，采用CHAID决策树算法进行智能分析，来达到提取犯罪嫌疑人人群特点的目的。邓灵评使用DBSCAN技术从大量犯罪数据中找出可能的犯罪团伙，但该方法在数据量巨大的时候会因为大量的计算而导致性能低下。

这些信用评分方法能在一定程度上实现电信诈骗团伙识别的目的，但在数据量巨大，特征丰富的条件下很难进一步提升识别的准确性与效率。对此，本发明在MR-DBSCAN簇类分析方法的基础上，采用GEOHASH算法来合理分配大数据计算的负担，并使用了spark并行计算框架来使得整个系统具有一个良好的扩展性，已达到在大数据条件下通过分布式计算来快速稳定的提取相关人员特征值，在多维度下进行簇类分析，并最终达到高效准确识别电信诈骗团伙的目的。

随着社会经济的快速发展，电信诈骗犯罪的发案率不断升高，相对于一般的刑事案件，电信诈骗犯罪一般为团伙作案，且可以借助方便的通信设备作案，对社会负面影响范围更广。如何遏制和打击电信诈骗犯罪团伙已经成为各地公安机关当前工作的重点任务。通过数据挖掘对相关人员进行聚类分析，可以有效地识别出犯罪团伙，该方法通常是从大量数据中挖掘出相关人员的特征信息，通过统计学技术识别出相关人员所处的犯罪团伙，从而帮助公安机关实现打击罪犯团伙的目的。但随着大数据时代的发展，人们的行为产生的数据特征和数据量快速增加，已经远远超出传统方法的数据处理能力。因此，传统的电信诈骗犯罪团伙识别方法并不能快速有效地在大数据环境下对犯罪团伙进行有效地甄别。为了更好的根据公安部门所掌握的数据来识别出电信诈骗犯罪团伙，创立一种支持多维度数据，高效，负载均衡的电信诈骗团伙甄别方法是很有必要的。  
近年来出现了很多针对电信诈骗团伙识别的研究。例如，潘芳等人提出了一种基于模糊层次聚类算法的犯罪网络分析方法；李亮利用快速分裂算法对已经建立的嫌疑人社会网络进行分解，来缩小可能的犯罪团伙排查范围。张玉超提出空间轨迹向量化方法，最终利用K-means聚类算法来实现人员聚类的目的。蔡霖翔在K-means聚类算法基础上，采用CHAID决策树算法进行智能分析，来达到提取犯罪嫌疑人人群特点的目的。邓灵评使用DBSCAN技术从大量犯罪数据中找出可能的犯罪团伙，但该方法在数据量巨大的时候会因为大量的计算而导致性能低下。  
这些信用评分方法能在一定程度上实现电信诈骗团伙识别的目的，但在数据量巨大，特征丰富的条件下很难进一步提升识别的准确性与效率。对此，本发明在MR-DBSCAN簇类分析方法的基础上，采用GEOHASH算法来合理分配大数据计算的负担，并使用了spark并行计算框架来使得整个系统具有一个良好的扩展性，已达到在大数据条件下通过分布式计算来快速稳定的提取相关人员特征值，在多维度下进行簇类分析，并最终达到高效准确识别电信诈骗团伙的目的。

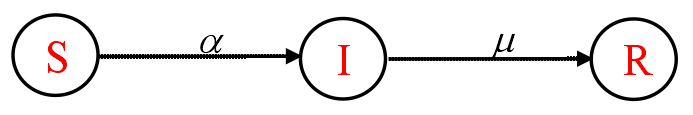
## 研究

电信诈骗团伙识别现状

瓶颈

我们的方法

疟疾是一种可以跨种群转播的疾病，可以通过水源、食物、身体接触等方式在宿主之间传播。在疾病的传播理论中基于传染病动力学一般分为 S、I、R 三个状态每个状态个体的特点如下：S 态（易感态）个体，该状态的个体不不具备传播传染病的能力，但是在传染病传播的过程中可能会被患病个体传染；I 态（感染态）个体，这类状态的个体已经患病，并且具有感染能力，可以将传染病传染给其他个体；R 态（恢复态）个体，这个状态的个体是已经被治愈并且具有对该种传染病免疫能力的个体，或者已经死亡的个体，这一类个体不会再被传染，同时也失去了传染性。



SIR动力学模型

疟疾的传播理论与DBSCAN算法有相似性，I状态和核心对象对应。类似于某个个体接触I状态的个体在一定概率上被转染，当一个和核心对象之间距离可达时，该点也可以划为类簇，不同的地方在于传染具备概率性而DBSCAN算法中不具有概率设定。与之类似的还有，S状态的个体和边界对象具备类似的属性。

在疾病的跨种群传播的研究中社会网络是常用的研究方法。社会网络中每个个体会被当成节点，个体之间的传播关系被当做边，边之间的连接有强弱之分，个体节点的作用有强弱之分。

// todo: 疾病传播的社会网络理论

疟疾在种群间的传播具备以下特性：如果一个个体P属于种群1，也属于种群2，如果个体为I状态的，则两个种群则都可能会被感染；如果个体为S状态的或R状态的则两个种群不会通过P进行种群建传播。

收到疟疾跨种群传播理论的启发，本文提出了并行化DBSCAN算法的交叠分区策略，并且提出了一种基于多维空间的GeoHash编码的分区策略，并且结合分布式图计算框架，提出类群合并算法。这些算法共同构成了并行化的DBSCAN算法。

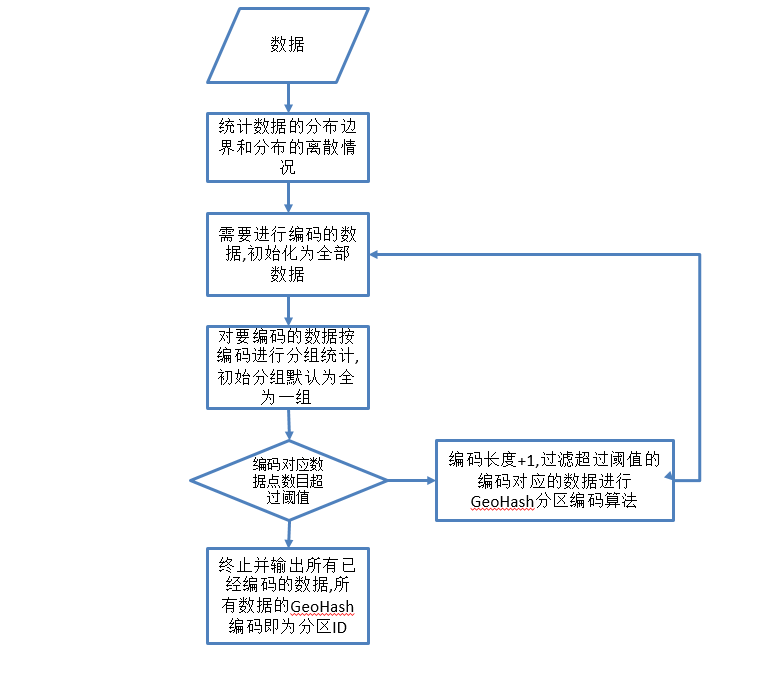
DBSCAN

GeoHash分区策略

交叠分区策略

基于分布式图计算的类群合并算法

1.整体算法



Todo:此处应该有一个shuffle的图

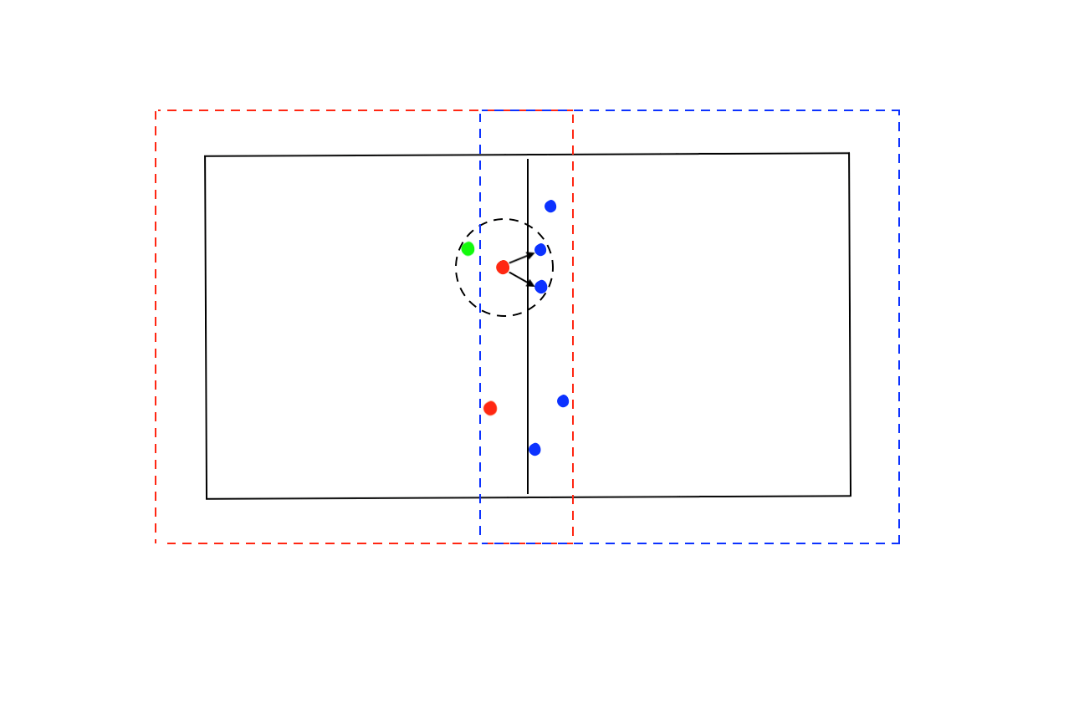
2.交叠分区策略

DBSCAN在进行并行化的过程中并行过程并不复杂，复杂的是两个相邻区域的DBSCAN运算结果如何合并。如下图:

受到疟疾在种群间的传播模型的启发，本文提出了DBSCAN在并行化时的定理规则:

两个相邻区域S1和S2,如果我们S1和S2延边界向外扩展epsilon个单位，则具有以下性质:

S1中的类簇如果P属于S1中C1类簇的核心点和S2.



3.GeoHash空间划分策略

定义如下:

1. 定义GeoHash编码函数,geohash(vector, loops), vector为要进行编码的向量，loops编码要进行的循环数
2. 定义考虑到交叠分区的GeoHash编码函数geoHashWithMirror(vector,loops),loops为编码要进行的循环数,vector为要进行编码的向量。
3. 最大分区点数目，如果超出该数目则进行分区否则不进行分区
4. 初始分区循环数，决定了开始时要对数据进行的geoHash循环数

geoHash算法:

输入:向量vector,循环次数loops

1. 初始化计数项count=1，编码结果bit=0，向量在所有维度的最小值边界为minBound，在所有维度的最大值边界为maxBound，minBound和maxBound为向量
2. 定义i为count对vector的长度取模，mean = (minBound[i] + maxBound[i]) / 2
3. 如果vector[i] <= mean:

minBound[i] = mean

bit = bit左移一位

否则:

maxBound[i] = mean

bit = bit左移一位 + 1

1. 重复第3步，每一步count+1直到count=loops时终止,输出bit

geoHashWithMirror算法:

输入:向量vector,循环次数loops

输出:向量的分区集合

1. 初始化计数项count=1，编码结果bit=0，向量在所有维度的最小值边界为minBound，在所有维度的最大值边界为maxBound，minBound和maxBound为向量
2. 定义i为count对vector的长度取模，mean = (minBound[i] + maxBound[i]) / 2
3. 如果|vector[i] – mean| < epsilon :

新建vectorMirror = vector, vectorMirror[i]= 2\*mean – vector[i]

对vectorMirror调用geoHash函数, geohash(vectorMirror,loops),将结果加入到分区集合中

否则:

进行第4步

1. 如果vector[i] <= mean:

minBound[i] = mean

bit = bit左移一位

否则:

maxBound[i] = mean

bit = bit左移一位 + 1

1. 重复3/4步，每一步count+1直到count=loops时终止,将bit加入到分区集合中

GeoHash分区策略

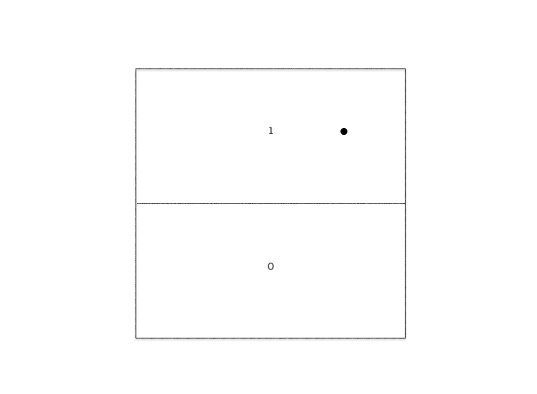
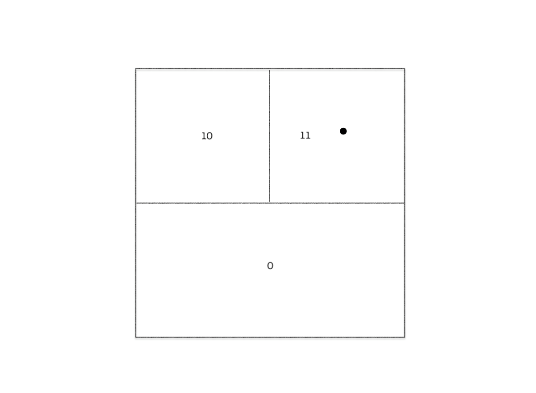
输入:数据，最大分区点数目，

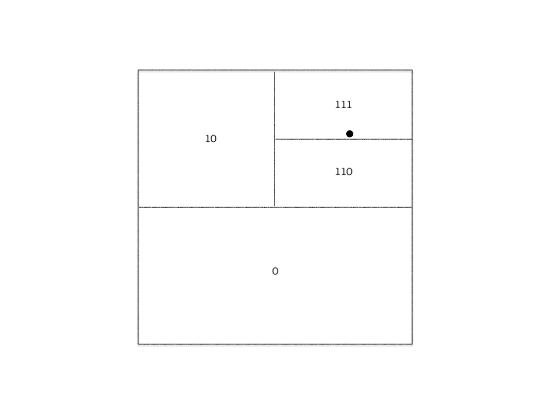
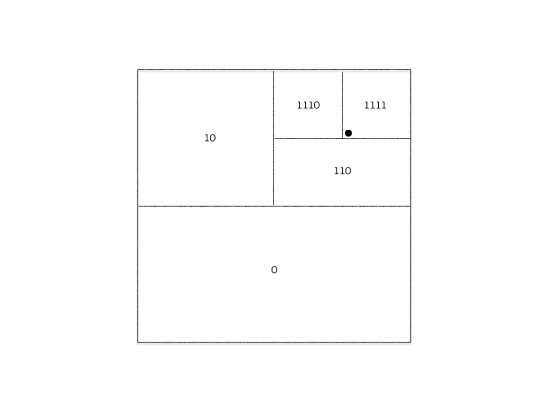
Step1.初始化所有的数据点,每个点对应一个分区集合,集合为空

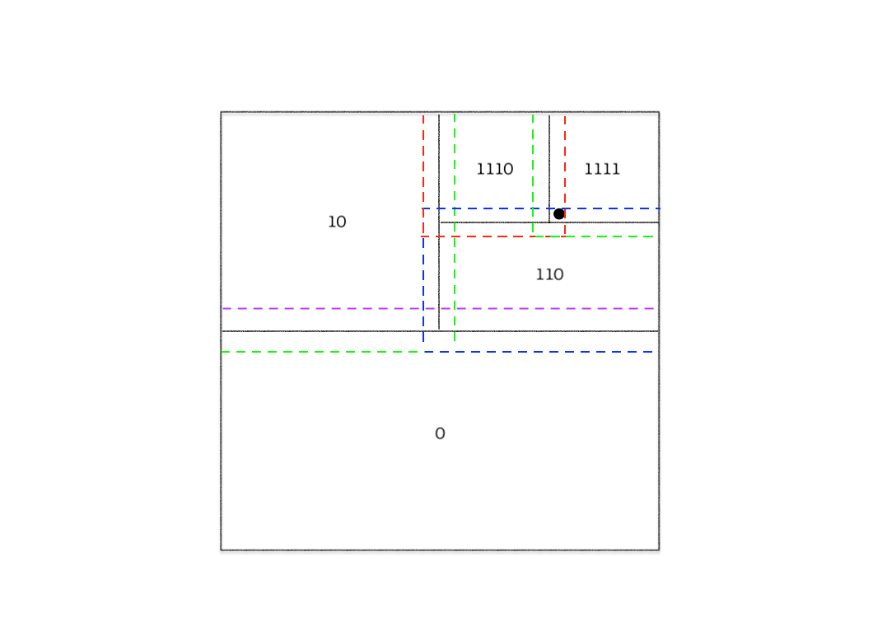
Step2.对所有的点调用geoHashWithMirror算法，以输出的编码集合为分区ID

Step3.统计每个分区ID对应的节点数,筛选数目超过最大分区点数目的分区ID，并筛选出其对应的数据

Step4.对3中筛选的数据调用Step2和Step3直至所有分区点的数目均满足条件。



4.分布式图计算的类群合并算法

定义:

1）分区Si中的类簇Cj的表示为Cj|Si,

2）如果某个点被划分到多个分区，则P称为这些分区之间的交叠点

DBSCAN的重要概念：

epsilon邻域,点数阈值minPoints

核心对象:如果点邻域内的点数目大于minPoints则认为该点为核心点

直接密度可达:如果xq在核心点xp的邻域内，则称xq到xp直接密度可达

密度可达：如果xp，xq

交叠分区类簇合并定理:

类似于跨种群传播理论中必须有一个节点为I状态同时连接两个种群.两个不同分区的类簇合并时也需要核心点进行传播。不同的是DBSCAN更加复杂，因为交叠点的邻域点可能在其他分区，所以交叠点在不同的分区可能有不同的状态。如在Si中为核心点，在Sj中则为噪声点.

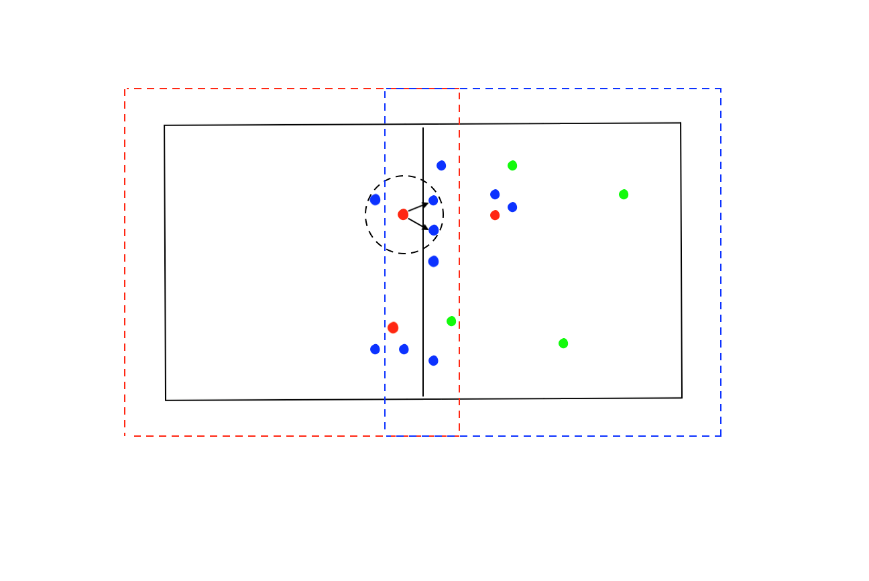
受到疟疾在种群间传播理论的启发，本文提出了DBSCAN在交叠分区中的类簇合并定理:

对于分区Si和Sj中的类簇Cu|Si和Cv|Sj，两个类簇为同一类簇的充分必要条件为:

至少存在一个交叠点P为两个分区的密度可达点，且P在至少其中一个分区为核心点.

现证明如下:

1. 首先证明充分条件,如果P在Si中为核心点，Sj中为密度可达点, 则在S1∪S2中P为核心点，又根据密度可达点的定义，Cv|Sj中必然存在核心点P’满足P’到P的距离小于epsilon。因此，任意Cu|Si中任意点P1，必然与P’密度相连，而Cv|Sj中的任意点P2必然与P密度相连，而P到P’直接密度可达，P’到P密度可达因此P1与P2密度相连，根据DBSCAN的算法定义，一个类簇为密度相连的最大集合，因此Cu|Si和Cv|Sj为同一类簇。
2. 再证明必要条件，如果两个类簇为同一类簇，根据DBSCAN对类簇的定义——密度相连的最大集合，按任意切分线将该类簇切分，则至少有一个核心点位于切分线的epsilon邻域中（否则切分线epsilon邻域内都没有点的话位于切分线两侧的点不可能有点使其密度相连，因为密度相连定义p -> p2 -> .. -> pi -> pi+1 -> q，由于所有点均位于分割线在epsilon的两侧所以必然有点pi在）。



最大联通子图

// Step1.统计所有分区内的Cj，在分区Pi内所有的Cj之间构建全联通图

Step1.初始化一个涉及到分区间合并的类簇的图，命名为G

Step2.遍历所有的交叠点，交叠点划分到的分区构成分区集合S = {S1,S2,..},

任意分区集合中的两两分区Si,Sj,如果P在Si和Sj有类别归属,且类别归属为

Cu和Cv，则：

以Cu|Si和Cv|Sj为顶点建立一条边并加入图G

Step2.对图G进行最大联通子图算法，并给每个联通子图一个标记

Step3.遍历所有的点:

如果点所属的类簇Cj|Si为G中的一个顶点,则以该类簇的联通子图标记作为类簇的标记.

否则类簇不变