基于傅里叶变换和连通图的聚类分析方法*

巨瑜芳，雷小锋，戴 斌，庄 伟，宋丰泰
（中国矿业大学 计算机科学与技术学院，江苏 徐州 221116）

摘 要：聚类是假设数据在具有某种群聚结构的前提下根据观察到的无标记的样本发现数据的最优划分。针对已有的聚类算法存在的缺点，假设数据样本的结果簇是密集的，且簇与簇之间区别明显，基于该假设提出一种基于傅里叶变换和连通图的聚类分析方法 FGClus。首先针对每个样本点计算 k 阶距离矩阵并序列化作为离散傅里叶变换的输入信号；然后抽取频域内幅值最小的复数项并构造输入序列进行傅里叶逆变换，得到在时域空间中的最佳阈值；最后利用该阈值结合连通图指导最终的聚类过程。实验表明，FGClus 算法克服了 K-means 算法聚类前需确定聚类个数、聚类结果对初始代表点的选取敏感、只能聚类球状数据等缺点，取得了良好的聚类效果。

关键词：聚类分析；离散傅里叶变换；连通图；最短路径 K 近邻查询；最佳阈值
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1001-3695(2012)08-2837-04
doi:10.3969/j.issn.1001-3695.2012.08.009

Cluster analysis methods based on Fourier transform and graph theory

JU Yu-fang, LEI Xiao-feng, DAI Bin, ZHUANG Wei, SONG Feng-tai
(*School of Computer Science & Technology, China University of Mining & Technology, Xuzhou Jiangsu* 221116, *China*)

Abstract: Clustering is to find the best partition of unlabeled observations under a certain group structure hypothesis. For the shortcomings in the existing clustering algorithms, this paper assumed that the results of the data sample was intensive and the differences among every cluster were significant. Based on the assumption it presented a cluster analysis method called FGClus based on discrete Fourier transform and graph theory. First, this method calculatd k-distance matrix of each sample point as a sequence of the input signal of discrete Fourier transform, then extracted the minimum amplitude of the complex frequency domain items and constructed the input sequence of inverse Fourier transform, to get the optimal threshold value of the space in the time domain. Finally, it used threshold and connected graph to guide the final clustering process. Large numbers of experiments show that FGClus algorithm can overcome existed shortcomings of K-means algorithm, such as the number of clusters must be determined before clustering, the results is sensitive on initial selection of representative points and it just can cluster spherical datas, which achieves good clustering results.

Key words: cluster analysis; discrete Fourier transform; connected graph; shortest path KNN query; optimal threshold

1 聚类结构假设及算法的提出

聚类分析作为统计学的一个分支和一种无监督的机器学习方法已有很长的研究历史。近十多年来随着数据挖掘的兴起而成为数据分析领域的一个研究热点。聚类分析在机器学习、数据挖掘、模式识别、生物学、统计学和化学等许多领域都得到了广泛的研究和应用。聚类分析可以作为一个独立的数据挖掘工具，用来获得对数据分布情况的了解。聚类就是把 n 维空间中的点应用某种方法对其进行分组，使得同一个簇中的对象有很大的相似度，而不同簇间的对象有很大的相异性。聚类源于很多领域，包括计算机、统计学、数学、生物学和经济学。聚类分析计算方法[1]主要有如下几种：分裂法（partitioning methods）、层次法（hierarchical methods）、基于密度的方法（density-based methods）、基于网格的方法（grid-based methods）、基于模型的方法（model-based methods）。

本文提出了一种基于傅里叶变换和图论的聚类分析方法，假设聚类结构一定要涵盖数据的分布结构，即假设的数据结构要有足够的表达能力。对于本算法，假设结果簇是密集的且簇与簇之间区别明显，FGClus 算法对符合该条件的数据集有较好的聚类结果。对于满足一定条件即具有较高相似度的样本可以划分为一个类簇。当任意两个样本点满足这种条件时就认为这两个样本点之间存在一条边，通过查找连通图得到聚类结果。

1.1 各聚类算法假设

为了找到一个通用性强且具有较高效率的聚类算法，人们从不同的角度出发提出了很多种聚类算法。在数据挖掘中最常用的算法有：K-means、CLARANS、BIRCH、CURE、DBSCAN、FCM 等算法。对于 K-means 聚类算法，当数据量特别大时，每

收稿日期：2011-12-30；**修回日期**：2012-02-27　　**基金项目**：江苏省基础研究计划资助项目（BK2009093）；中国矿业大学科技基金资助项目（OD080313）

作者简介：巨瑜芳（1989-），女，硕士研究生，主要研究方向为数据挖掘（ju_yufang@126.com）；雷小锋（1975-），男，副教授，硕导，博士（后），主要研究方向为数据库、数据挖掘、机器学习图像处理等；戴斌（1982-），男，硕士研究生，主要研究方向为数据挖掘、机器学习；庄伟（1986-），男，硕士研究生，主要研究方向为图像处理、机器学习等；宋丰泰（1989-），男，硕士研究生，主要研究方向为数据挖掘、模式识别等.

次迭代都必须重新计算样本点到聚类中心的距离,具有较大的时间复杂度,算法计算效率低且不稳定。其次是领域知识依赖性大(聚类之前必须给出聚类数目和初始聚类中心),不同的初始聚类中心可能导致不同的聚类结果,并且会陷入局部最优,而且它只适合聚类大小差别不大的球状类簇[2]。CLARANS 算法也会陷入局部最优,虽然对噪声数据不敏感但是对数据的输入顺序比较敏感,同 K-means 算法一样,它只能聚类凸状或者球形的数据。BIRCH 算法核心是采用了一个三元组的聚类特征树汇总一个簇的有关信息,它也只适合于聚类凸状或球状的情况,需要输入正确的聚类数。CURE 选择数据空间中的固定数目具有代表性的点来代表各个簇,选择多个代表使得该算法可以聚类任意形状的类,簇的收缩和凝聚可以有助于控制噪声对结果产生的影响[3]。

本文主要针对 K-means 算法存在的缺点,如聚类之前必须确定聚类个数,只能聚类球状而且大小差别不大的类簇结构等, 提出了基于傅里叶变换和图论的 FGClus 算法,该算法在聚类之前不需要输入聚类个数和初始聚类中心,可以聚类任意大小任意形状的数据集。实验结果证明本算法对结果簇是密集的,且簇与簇之间区别明显时,能得到较高质量的聚类结果和最佳的聚类数目。

类簇定义如下:对于样本集 $U=\{x[n]\}(0\leqslant n<N;x[i],x[j]\in U)$,如果 $d(x[i],x[j])<\text{dis}$,则 $x[i],x[j]\in C_k$。其中:$x[i],x[j]$表示任意两个样本点, $d(x[i],x[j])$为两个样本点之间的距离,当该距离小于阈值 dis 时,表示样本点 $x[i]$和 $x[j]$具有较高的相似性,应归为一类; C_k 表示第 k 个类。

1.2 FGClus 算法的提出

本算法假设结果簇是密集的,簇与簇之间区别明显,因此需要找到某个距离使得大多数距离值都在该值附近,为此引入傅里叶变换,使用傅里叶变换来得到该最佳距离值,用该值指导聚类。FGClus 算法的提出主要利用了离散傅里叶变换和连通图的方法。

离散傅里叶变换(discrete fourier transform,DFT)在时域和频域上都呈离散的形式,将信号的时域采样变换为其 DTFT 的频域采样。在形式上,变换两端(时域和频域上)的序列是有限长的,而实际上这两组序列都应当被认为是离散周期信号的主值序列。即使对有限长的离散信号作 DFT,也应当将其看做其周期延拓的变换[4]。

对于 N 点序列 $\{x[n]\}(0\leqslant n<N)$,它的离散傅里叶变换(DFT)为

$$X[k]=\sum_{n=0}^{N-1}e^{-i\frac{2\pi}{N}kn}x[n]\quad k=0,1,2,\cdots,N-1 \tag{1}$$

也可以用正余弦叠加的方法表示傅里叶变换。通过把输入信号和每一种频率的正余弦信号进行相乘(关联操作),从而得到原始信号与每种频率的关联程度(即总和大小),具体的计算方法如下:

$$\begin{cases}\text{Re}\ X[k]=\sum_{n=0}^{N-1}x[n]\cos(2\pi kn/N)\\ \text{Im}\ X[k]=-\sum_{n=0}^{N-1}x[n]\sin(2\pi kn/N)\end{cases}\quad k=0,1,2,\cdots,N-1 \tag{2}$$

其中:频谱的振幅大小为

$$\sqrt{\text{Re}\ X[k]^2+\text{Im}\ X[k]^2}$$

它与频率一一对应的。

离散傅里叶变换的逆变换(IDFT)为

$$x[n]=\frac{1}{N}\sum_{k=0}^{N-1}e^{i\frac{2\pi}{N}nk}X[k]\quad n=0,1,2,\cdots,N-1 \tag{3}$$

在 FGClus 算法中,把求出符合要求的距离序列作为时间离散的信号序列,通过离散傅里叶变换式(2)计算得到对应的复数序列。对该序列的每一项进行求模运算,得到一组实数序列,这组实数序列与信号含有的各次谐波的频率是一一对应的,即为某一谐波的幅值。幅值的大小反映频率的大小。找到振幅最小时的傅里叶变换值,把它作为傅里叶逆变换输入序列 $X[k]$的某一序列项,其余项置 0;然后再对该序列进行傅里叶逆变换得到对该频率贡献最大的一个时间序列值,那么这个值即为找到的距离阈值 dis;判断点与点之间的距离是否小于 dis,如果小于则认为有边存在,通过查找连通图得到聚类。

FGClus 算法将样本中点与点的距离抽象为无向图来处理。该无向图由顶点集 V 和边集 E 组成,记为 $G=(V,E)$,其中顶点集为样本点,点与点之间的距离小于等于 dis 时就认为有边存在,置为 1,否则就认为没有边,置为 0,即构成一个 0-1 的二值矩阵。

在该无向图中,若从顶点 u 到顶点 v 有路径,则 u 和 v 是连通的,运用傅里叶变换计算出的最佳距离 dis 来确定样本点的连通性,利用深度优先遍历方法查找连通图的个数,即为最后聚类的数目。

2 算法描述和理论分析

本算法的主要思想旨在找出两两样本点之间距离出现在某个范围之内(该范围内的距离相差很小)的最佳距离,然后通过该距离指导聚类。怎样找到某一范围内距离相差很小的样本点呢? 如果将这些距离映射到频域内,那么距离相差很小时频率的变换幅度越小。自然而然想到傅里叶变换,它是数字信号处理领域一种很重要的算法。傅里叶原理表明:任何测量到的时序或信号,都可以表示为不同频率的正弦波信号的叠加。与傅里叶变换算法对应的傅里叶逆变换算法从本质上说也是一种累加处理,这样就可以将单独改变的正弦波信号转换成一个信号。因此可以说,傅里叶变换将原来难以处理的时域信号转换成了易于分析的频域信号,可以利用一些工具对这些频域信号进行处理、加工。最后再利用傅里叶逆变换将这些频域信号转换成时域信号。

对于图像的傅里叶变换得到的频率表征了图像中灰度变换剧烈程度的指标,是灰度在平面空间上的梯度。如:大面积的沙漠在图像中是一片灰度变化缓慢的区域,对应的频率值很低;而对于地表属性变换剧烈的边缘区域在图像中是一片灰度变化剧烈的区域,对应的频率值较高。傅里叶变换在实际中有非常明显的物理意义,设 f 是一个能量有限的模拟信号,则其傅里叶变换就表示 f 的谱。在 FGClus 算法中, 首先计算所有点到每一个点的距离,为了得到比较精确的聚类结果,取距离每一个样本点最近的 k 个距离构造傅里叶输入序列。同图像的傅里叶变换类似,对距离数组进行傅里叶变换得到的频率表征了样本点之间距离的变化程度,因此低频值表征了各个距离值相差很小即样本点的密度较大,这样可以筛选出样本中密度

较大的样本点;然后使用该最低频率并将其余序列项置零构造傅里叶逆变换的输入序列。逆变换得到一组时序序列,则该时序序列中最大值即为对输入的最低频率贡献最大的值,这个值就是两两样本点之间距离出现在某个范围之内(该范围内的距离相差很小)的最佳距离。

FGClus 算法的整体思想为:对每一个样本点计算所有其他样本点到该样本点的距离,然后取离每一个样本点最近的 k 个距离作为初始信号序列,输入离散傅里叶变换,计算得到该信号序列所对应的频域复数序列,利用该序列中的最小振幅复数项构造离散傅里叶逆变换的输入序列,最后利用得到的逆变换值作为阈值距离 dis 来指导聚类。

具体算法描述如下:

输入:样本集 $D=(d_1,d_2,\cdots,d_m)$,包含 m 个样本点,每个样本点有 n 个属性;取距离每个点最近的 k 个点,点与点之间采用欧氏距离计算。

输出:聚类类别数及聚类的结果。

a)对样本集 D 中的每一个样本点计算所有其他样本点到该点的距离,构造距离矩阵 $T_{m\times m}$。

$$d(i,j)=d(j,i)=\begin{cases}\text{dist}_{(i,j)} & i\neq j\\ 0 & i=j\end{cases} \tag{4}$$

$$\text{即}\quad T=\begin{bmatrix}0 & d_{(1,2)} & d_{(1,3)} & \cdots & d_{(1,m)}\\ d_{(2,1)} & 0 & d_{(2,3)} & \cdots & d_{(2,m)}\\ d_{(3,1)} & d_{(3,2)} & 0 & \cdots & d_{(3,m)}\\ \vdots & \vdots & \vdots & & \vdots\\ d_{(m,1)} & d_{(m,2)} & d_{(m,3)} & \cdots & 0\end{bmatrix}$$

b)对距离矩阵 $T_{m\times m}$ 的每行元素进行升序排列,取距每个样本点最小的 k 个距离,组成长度为 $m\times k$ 的一维序列。

c)对步骤 b)的一维序列进行离散傅里叶变换,得到对应频域内的复数序列,取其中振幅最小时对应的复数项。

d)构造离散傅里叶逆变换输入序列。针对步骤 c)得到的傅里叶变换序列,保持最小振幅所对应的复数项不变,其余序列项的实部和虚部分别置零。

e)运用该距离判断样本点与点之间是否有边,查找连通图数量并得到聚类个数和聚类结果。

f)改变 k 值直到在某一范围内聚类结果保持不变,而在范围之外改变剧烈,那么 k 在这个范围内时聚类效果最佳。

g)处理孤立点。查找离孤立点最近的点所在的类进行归并。

h)输出聚类结果和聚类数。

FGClus 算法需要两个数据结构:二维数组 $d[m][m]$ 和队列 queue,$d[m][m]$ 存储每个样本点到其他所有样本点的距离,队列 queue 用于存储深度遍历过程中查找到的符合条件的样本点,FGClus 算法的时间复杂度为 $O(n^2)$。

3　仿真实验

3.1　环境设置和实验设计

实验的硬件环境为英特尔双核 2.0 GHz 处理器和 1 GB 内存,软件平台为 Windows XP 操作系统和 Eclipse 编程平台,使用 Java 语言编程实现。

实验对 m 个 n 维数据进行聚类,k 是距离每个点最近的 k 个距离用以构造傅里叶变换序列。实验主要分为三部分,首先通过多组模拟数据验证算法的有效性,再根据实验结果确定 k 的取值,最后分别使用 K-means 算法和 FGClus 算法对真实数据集进行聚类分析,验证 FGClus 算法能得到样本的最佳聚类类别数和较好的聚类结果。

3.2　模拟数据实验

实验模拟生成了六组数据,由于篇幅限制仅给出 data1 和 data2 的实验结果,直接通过视觉判断分析聚类结果的有效性。当 k 的取值在一定范围内并达到聚类稳定时得到聚类个数 p,然后采用 K-means 算法对这两组数据聚为 p 类,对比两个算法的聚类结果。

对于 data1,当 k 的取值在 10 ~ 15 时,聚类结果稳定,聚类个数为 3,结果如图 1 所示;然后对 data1 采用 K-means 算法聚为 3 类,聚类结果如图 2 所示。

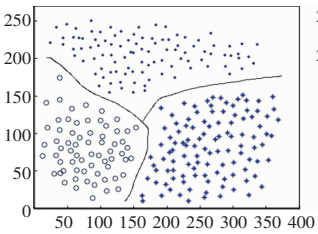

图 1　FGClus 算法的聚类结果

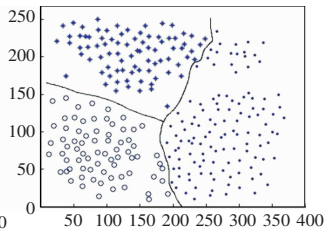

图 2　K-means 算法的聚类结果

对于 data2,当 k 的取值在 9 ~ 15 时,聚类结果稳定,聚类个数为 4,结果如图 3 所示;同样对 data2 采用 K-means 算法聚为 4 类,聚类结果如图 4 所示。

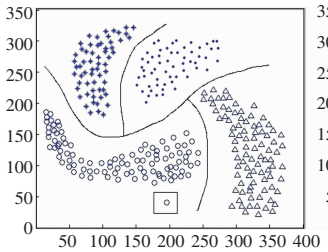

图 3　FGClus 算法的聚类结果

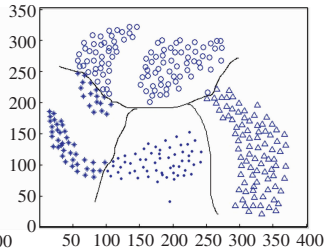

图 4　K-means 算法的聚类结果

FGClus 算法的时间复杂度为 $O(n^2)$,标准 K-means 算法的时间复杂度为 $O(\log nkt)$,n 为样本点的个数,k 为聚类个数,t 为迭代次数。两个算法分别对两组数据的耗时如表 1 所示。

表 1　K-means 和 FGClus 算法对两组测试数据的耗时对比

数据	K-means 算法/s	FGClus 算法/s
data1	0.058 7	0.061 3
data2	0.045 8	0.049 2

从两组实验结果可以看出,本算法能较好地给出聚类类别数,并得到较为满意的聚类结果。对于 data2,本算法能找出孤立点并对孤立点进行归并处理,因此孤立点不会对结果产生影响。从两组数据的分布情况和两个算法分别对两组数据聚类得到的结果可以看出,FGClus 算法可以发现任意形状的类簇并克服了 K-means 算法在聚类之前必须给出聚类数目和初始聚类中心(不同的初始聚类中心点导致不同的聚类结果)的缺点。从表 1 可以看出,FGClus 算法相对于 K-means 算法耗费了较多的执行时间,当样本数据量非常大时,FGClus 算法的执行效率较低。因此,对于 FGClus 算法的时间复杂度有待进一步研究改进。通过实验证明该算法的聚类结果具有较高的准确

率且聚类结果是稳定的。

然而已有很多学者关于 K-means 算法提出了各种不同的改进方法,典型的改进方法例如:

a)对初始中心点的选取方法的改进,通过计算每个数据对象的密度参数,然后选取 k 个处于高密度分布的点作为初始聚类中心点[5];

b)针对每次调整簇中心后确定新的簇中心需要大量的距离计算,提出一种利用簇中心的变换信息来确定新簇中心的方法,从而减少了过滤算法的计算复杂度[6]。

这些方法都是利用优化初始点的选择或者降低计算法复杂度为出发点进行改进的,但是传统的 K-means 算法对初值的依赖性很强,无论怎么样优化初始点的选择或计算法复杂度都需要事先给出要生成的簇的数目 k,而这个参数 k 的确定一般是根据用户的经验知识给出的;另外,初始中心点的选取具有很大的随机性,这种随机性往往导致了聚类结果的不稳定性。FGClus 算法主要克服了这个缺点,无须人为给出聚类数目,只需要通过傅里叶变换得到适当的阈值,然后通过该阈值查找连通图的个数即为最后聚类的数目。从实验结果可以看出该算法能够得到比较准确的聚类类别数和稳定的聚类结果。

3.3 参数 k 的确定

本实验模拟生成三组数据,这三组数据的簇间相似度尽可能低,簇内相似度高,然后通过聚类结果确定 k 的取值(表 2)。

表 2 实验确定 k 的取值

数据	样本个数	聚类稳定时的类别数	聚类稳定时 k 的取值
第一组	50	3	7 ~ 13
第二组	100	3	8 ~ 15
第三组	200	4	6 ~ 13

如图 5 所示,横纵坐标分别表示 k 的取值和分类类别数。三组实验表明,当 k 的取值在一定范围内时,聚类类别和聚类结果达到稳定值,说明这个时候的聚类个数最佳。大量实验表明,k 的取值通常在 6 ~ 15 之间时,得到了最好的聚类结果。

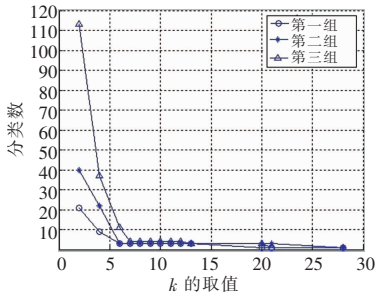

图 5 k 的取值与分类数的关系曲线

3.4 真实数据实验

本实验使用 K-means 算法和 FGClus 算法对从《中文文本分类语料库-TanCorpV1.0》中选取的财经、科技、体育三类各 40 篇文章计算得到的 120 个 20 维的特征向量聚类,根据聚类结果的准确率和召回率来衡量算法的有效性。准确率 P 和召回率 R 度量公式定义[7]如下:

$$P_i = \frac{TP_i}{TP_i + FP_i}, R_i = \frac{TP_i}{TP_i + FN_i} \tag{5}$$

其中:i 表示第 i 个类;TP_i 表示算法认定属于第 i 类且确实属于第 i 类的样本数;FP_i 表示算法认定属于第 i 类而实际上不属于第 i 类的样本数;FN_i 表示算法认定不属于第 i 类而实际上属于第 i 类的样本数。最后的实验结果如表 3 所示。

表 3 分别使用 K-means 和 FGClus 算法对 120 个样本聚类结果的比较

算法	准确率(P)比较		
	财经/%	科技/%	体育/%
K-means	71.43	68.63	73.17
FGClus	85.71	81.82	80.49
算法	召回率(R)比较		
	财经/%	科技/%	体育/%
K-means	50.00	87.50	75.00
FGClus	75.00	90.00	82.50

使用 FGClus 算法对 120 个样本进行聚类,聚类结果稳定时聚类类别数和标准聚类结果一样,也分为三类。从表 3 中可以看出本算法在聚类结果的准确率和召回率上都优于K-means 算法。

4 结束语

本文主要针对 K-means 聚类算法存在的缺点提出一种基于傅里叶和图论的聚类分析方法,在聚类时无须事先知道聚类个数,可以聚类大小差别很大的类簇和任意形状的符合簇内相似度高、簇间相似度低的数据集,实验表明该方法能够得到聚类稳定时的最佳聚类个数和较好的聚类结果。该算法在结果的有效性上具有很大的提高,然而时间复杂度较大,这是下一步改进的重点。

参考文献:

[1] HAN Jia-wei, KAMBER M. Data mining: concepts and techniques [M]. San Francisco: Morgan Kaufmann Publishers, 2011: 223-250.

[2] SHI Na, LIU Xu-min. Research on K-means clustering algorithm: an improved K-means clustering algorithm [C]//Proc of the 3rd International Symposium on Intelligent Information Technology and Security Informatics. 2010: 1-3.

[3] 牟廉明. 数据挖掘中聚类方法比较研究[J]. 内江师范学院学报, 2003, 16(10): 3-4.

[4] 程乾生. 数字信号处理[M]. 北京: 北京大学出版社, 2010: 155-162.

[5] 韩凌波, 王强, 蒋正峰, 等. 一种改进的 K-means 初始聚类中心选取算法[J]. 计算机工程与应用, 2010, 46(17): 150-152.

[6] 安建成, 史德增. 一种改进的 K-means 算法[J]. 电脑开发与应用, 2011, 23(8): 31-33, 60.

[7] 雷小峰, 杨阳, 张克, 等. 一种基于元启发式策略的迭代自学习 K-means 算法[J]. 计算机科学, 2009, 36(7): 175-178.

[8] CORMEN T H, LENISERSON C E, RIVEST R L, *et al*. 算法导论[M]. 潘金贵, 顾铁成, 李成法, 等译. 北京. 机械工业出版社, 2006: 322-335.

[9] HAN Jia-wei, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 等译. 北京. 机械工业出版社, 2007.

[10] 徐岩, 张晓明, 王瑜, 等. 基于离散傅里叶变换的频谱分析新方法[J]. 电力系统保护与控制, 2011, 39(11): 38-43.

[11] LIU Hong. Internet public opinion hotspot detection and analysis based on K-means and SVM algorithm [C]//Proc of International Conference on Information Science and Management Engineering. 2010: 257-261.

[12] MOSHE S, HERTZ D. On computing DFT of real *N*-point vector and IDFT of DFT-transformed real N-point vector via single DFT [J]. IEEE Signal Processing Letters, 1999, 6(6): 141.