山东农业大学学报(自然科学版),2017,48(1):100-103 VOL.48 NO.1 2017
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2017.01.020

基于聚类分析的数据挖掘方法研究

黄 蓉
湖南科技职业学院, 湖南 长沙 410007

摘 要：针对FCM算法的聚类效果易受其初始中心位置影响和易陷入局部最优的缺点，将灰狼优化算法和FCM结合，提出一种基于GWO优化FCM的聚类分析方法。以KDD CUP99数据集为研究对象，研究结果表明，与PSO、GA和SA算法相比较，GWO算法聚类分析的准确率和误判率更低，具有更快的收敛速度，效果更优，从而为数据聚类分析提供新的方法和途径。

关键词：聚类分析; 数据挖掘
中图法分类号：TP391.1 **文献标识码**：A **文章编号**：1000-2324(2017)01-0100-04

Study on Data Mining Based on Clustering Analysis

HUANG Rong
Hunan Vocational College of Science and Technology, Changsha 410007, *China*

Abstract: Aiming to the clustering effect on FCM algorithm is easily affected by the initial location and fell into local optima, the gray optimization algorithm is applied to improving FCM to put forward a method of clustering analysis GWO_FCM. The KDD CUP99 data set as the research object, the research results showed that GWO algorithm of clustering had a accuracy and lower error rate, comparing with PSO, GA and SA algorithm, it was with faster convergence speed and better effect so as to provide a new approach for data clustering analysis.

Keywords: Clustering Analysis; data mining

聚类分析是数据挖掘技术中一项重要技术，其根据一定规则将数据集分成若干组或者若干类的过程，同组或者同类数据具有一定相似度，不同组或不同类中的数据在划分规则上不具相似性。目前，聚类分析被广泛地应用于模式识别、图像处理、市场分析和数据分析等领域。目前聚类理论较为完善的算法为FCM算法和K-means算法。FCM算法[1]具有算法简单、收敛速度快、处理大数据集等优点，但其聚类结果易受初始中心位置影响，且结果容易陷入局部极小值。

针对FCM算法易受初始中心点位置和陷入局部最优的缺点，将灰狼优化算法[2]（Grey Wolf Optimization，GWO）和FCM算法结合，提出一种GWO优化FCM的聚类分析方法，避免FCM算法陷入局部最优。

1 FCM 算法

FCM聚类分析的目标函数如公式（1）所示[3]：$J_{\mathrm{m}}(U,Z)=\sum_{j=1}^{N}\sum_{i=1}^{C}\left(u_{ij}\right)^{m}d^{2}\left(\mathbf{x}_{j},\mathbf{z}_{i}\right)$ (1)

若 p 表示样本 X_j 的维数，则 $X=\{X_1,X_2,...X_j,...X_N\}$ 表示为 $p\times N$ 的矩阵；N 表示样本数目；C 表示聚类数目；$u_{ij}\subseteq \mathrm{U}(p\times N\times C)$ 表示矢量 X_j 隶属于第 i 类的隶属度函数，满足 $u_{ij}\in[0,1]$ 且 $\sum_{i=1}^{C}u_{ij}=1,\left(j=1,2,...,N\right)$；聚类中心 $Z=\{Z_1,Z_2,...,Z_i,...Z_C\}$ 表示 $p\times C$ 的矩阵，u_{ij} 和 Z_i 根据公式（2）进行聚类迭代更新：

$$\begin{cases}u_{ij}=\dfrac{1}{\sum_{k=1}^{C}\left(d\left(\mathbf{x}_{j},\mathbf{z}_{i}\right)/d\left(\mathbf{x}_{j},\mathbf{z}_{k}\right)\right)^{2/(m-1)}}\\ \mathbf{z}_{i}=\sum_{j=1}^{N}\left(u_{ij}\right)^{m}\mathbf{x}_{j}\Big/\sum_{j=1}^{N}\left(u_{ij}\right)^{m},\left(i=1,2,...C\right)\end{cases}\tag{2}$$

对于每个模糊隶属度，由 $m\in(1,\infty)$ 控制模糊度的权重指数；$d^2(X_j,Z_i)=\|X_j-Z_i\|$ 表示相似性测度。P 表示数据样本数据的维数；N 表示数据点数目；X_i 表示数据第 i 个特征；C 表示聚类类别数；u_{ij} 数据点 i 属于第 j 类的隶属度；Z_i 表示第 i 类聚类中心。FCM算法聚类过程如下：

收稿日期: 2016-11-12 **修回日期**: 2016-12-16
基金项目: 2014年湖南省教育厅科学研究项目:基于用户兴趣挖掘技术的移动校园信息推送系统研究及应用（14C0505）; 湖南科技职业学院重点科研课题:2013年基于移动终端的个性化校园信息推送系统的研究与实现(KJ13102)
作者简介: 黄 蓉(1981-),女,硕士,讲师,主要研究方向为数据挖掘、职业教育. E-mail:ocean1205@163.com

数字优先出版:2017-02-17 http://www.cnki.net

Step 1：确定 C 和 m，设定迭代停止阈值ε>0，置迭代次数 t=0，初始化聚类中心 Z。

Step 2：计算隶属度矩阵：$$u_{ij}=\frac{1}{\sum_{k=1}^{C}\left(d\left(\mathbf{x}_j,\mathbf{z}_i\right)/d\left(\mathbf{x}_j,\mathbf{z}_k\right)\right)^{2/(m-1)}} \tag{3}$$

Step 3：计算新的聚类中心：$$\mathbf{z}_i=\sum_{j=1}^{N}\left(u_{ij}\right)^m\mathbf{x}_j\Big/\sum_{j=1}^{N}\left(u_{ij}\right)^m,(i=1,2,...C) \tag{4}$$

Step 4：若$\left\| Z(k+1)-Z(k)\right\|$<$\varepsilon$停止，否则 $k=k+1$ 转 Step 2。

2　GWO 优化算法

灰狼优化算法（Grey Wolf Optimizer，GWO）是受灰狼觅食行为启发而提出的群智能算法，在 GWO 算法中，灰狼种群被划分成为α、β、δ、ω四类。优化过程中，狼群根据公式（1）和公式（2）更新种群位置[4,5]，其位置更新机制如图 1 所示，处于位置（X，Y）的灰狼可根据其附近猎物位置进行重新定位。

$$\vec{D}=|\vec{C}\cdot\vec{X}_p(t)-\vec{X}(t)| \tag{5}$$

$$\vec{X}(t+1)=\vec{X}_p(t)-\vec{A}\cdot\vec{D} \tag{6}$$

其中，$\vec{A}=2a\cdot\vec{r}_1-a$，$\vec{C}=2\cdot\vec{r}_2$，$t$ 表示当前迭代次数，$\vec{X}_p$ 表示猎物的位置向量，$\vec{X}$ 表示灰狼的位置向量，在迭代过程中，α是从 2 到 0 线性下降的，r_1，r_2 是处于[0,1]的随机向量。

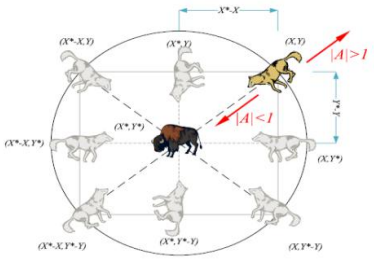

图 1　狼群位置更新机制

Fig.1 Location update mechanism for wolves

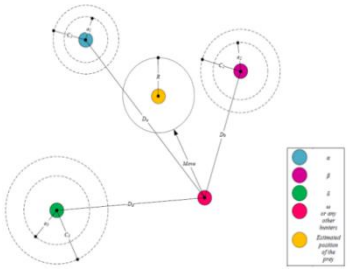

图 2 GWO 算法位置更新

Fig.2 Location update mechanism for GWO

图 1 中，虽然灰狼有 8 个可能位置，随机参数 A 和 C 允许灰狼转移到任何位置在周围猎物附近的连续的空间内。在 GWO 算法中，总是假定α、β、δ 是猎物最可能的位置（最优位置）。在优化过程中，到目前为止的前三个变量的最优解分别被设定为α、β、δ 。之后，其他灰狼被看成ω，能够根据α、β、δ 进行重新定位。重新调整狼群ω的位置的数学模型如下[5]：

$$\vec{D}_\alpha=|\vec{C}_1\cdot\vec{X}_\alpha-\vec{X}| \tag{7}$$

$$\vec{D}_\beta=|\vec{C}_2\cdot\vec{X}_\beta-\vec{X}| \tag{8}$$

$$\vec{D}_\delta=|\vec{C}_3\cdot\vec{X}_\delta-\vec{X}| \tag{9}$$

其中，$\vec{X}_\alpha$、$\vec{X}_\beta$、$\vec{X}_\delta$ 分别表示α、β、δ 的位置，$\vec{C}_1$、$\vec{C}_2$、$\vec{C}_3$ 分别表示随机向量，$\vec{X}$ 表示当前解的位置。根据公式（7）、公式（8）和公式（9）分别计算当前解和α、β、δ 之间的近似距离。定义距离之后，当前解的最终位置计算如下[5]：

$$\vec{X}_1=\vec{X}_\alpha-\vec{A}_1\cdot(\vec{D}_\alpha) \tag{10}$$

$$\vec{X}_2=\vec{X}_\beta-\vec{A}_2\cdot(\vec{D}_\beta) \tag{11}$$

$$\vec{X}_3=\vec{X}_\delta-\vec{A}_3\cdot(\vec{D}_\delta) \tag{12}$$

$$\vec{X}(t+1)=\frac{\vec{X}_1+\vec{X}_2+\vec{X}_3}{3} \tag{13}$$

其中，$\vec{X}_\alpha$、$\vec{X}_\beta$、$\vec{X}_\delta$ 分别表示α、β、δ 的位置，$\vec{A}_1$、$\vec{A}_2$、$\vec{A}_3$ 表示随机向量，t 表示迭代次数。图 2 表示 GWO 算法的位置更新原理图。公式（7）、（8）和（9）定义狼群ω分别到狼群α、β、δ 的步长，公式（10）、（11）、（12）和（13）定义狼群ω的最终位置。在 GWO 算法中，有两个随机向量分别为 $\vec{A}$、$\vec{C}$，这两个随机和自适应向量为 GWO 算法提供搜索和寻优的依据。图 1 中，若$|A|<1$，

狼群向猎物方向搜索；若$|A|>1$，狼群将偏离猎物，去寻找一个更适合的猎物。GWO 算法中的另一个随机变量$\vec{C}$，在公式（7）、（8）和（9）中，向量$\vec{C}$处于$[0,2]$。该向量为猎物提供随机权重，根据公式（1）随机加强($|C|>1$)或者随机削弱($|C|<1$)猎物的影响。GWO 算法流程如下：

Step 1：依据变量的上限和下限，随机初始化产生种群；

Step 2：计算每个灰狼的目标函数值（适应度函数值）；

Step 3：选择前三个最优灰狼，并将其保存为α、β、δ；

Step 4：根据公式（7）-公式（13），更新其他狼群（狼群ω）的位置；

Step 5：更新参数α、A、C；

Step 6：若不满足停止准则，则返回 Step 2；

Step 7：返回α的位置作为最优近似解。

3　基于 GWO 算法优化 FCM 算法

为避免 FCM 陷入局部最优和最优聚类初始中心点的选择，适应度函数如公式（14）所示[6]：

$$Minimize\ Fitness(C,m)=\frac{k}{J_m(U,Z)} \tag{14}$$

其中，k 表示常数，$J_m(U,Z)$ 表示总的类间离散度和。若聚类效果越好，则 $J_m(U,Z)$ 越小，个体适应度 $Fitness(C,m)$ 就越高。GWO 优化 FCM 算法聚类流程如下：

Step 1：GWO 算法参数设置，种群规模 N 和最大迭代次数 $Iteration$，根据搜索空间，随机产生 a、A、C；

Step 2：依据变量的上限和下限，随机初始化产生种群；

Step 3：计算每个灰狼个体适应度，并进行排序；

Step 4：根据适应度大小，选择前三个最优灰狼个体，并将其保存为α、β、δ；

Step 5：根据公式（7）-公式（13），更新其他狼群（狼群ω）的位置；

Step 6：更新参数α、A、C；

Step 7：判定算法是否满足停止准则。若满足，则终止，获得最优聚类中心；否则，返回 Step 3；

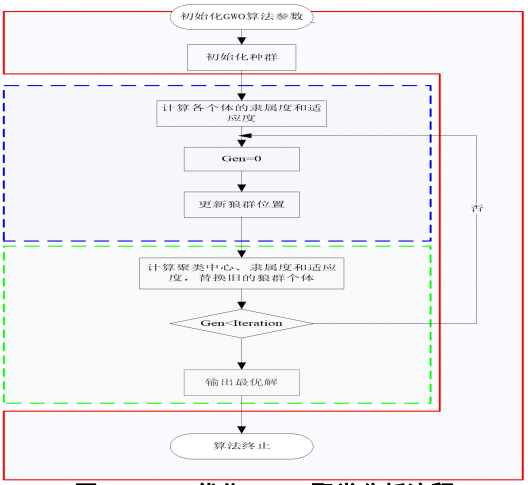

图 3 GWO 优化 FCM 聚类分析流程

Fig.3 Clustering Analysis Process based on GWO_FCM

4　实验分析

4.1　数据来源

为了验证 GWO 算法优化 FCM 聚类分析的有效性和可靠性，计算机操作系统为 Windows7，选择 MATLAB2015（a）软件为聚类数据分析平台，内存 4GB、处理器 P4 2.4GHZ。以 KDD CUP99 数据集为研究对象[7]，验证算法的有效性。该数据集每一组记录数据包括 41 个特征属性。

4.2　评价指标

为验证聚类分析的效果，选择准确率 $Accuracy$ 和误判率 $False_{ij}$ [7,8]为聚类分析效果的评价指标。

（1）准确率 $Accuracy$：若正确判断的类型数目为 a，实际的类型数目为 b，数据类型判断的准确率为：　$Accuracy = \frac{a}{b} \times 100\%$　(15)

（2）误判率 $False_{ij}$：若 a 为第 i 类数据类型的实际数目，c 表示将第 i 类数据类型误判为第 j 类数据类型的数目，则数据类型判断的误判率为：　$False_{ij} = \frac{c}{a} \times 100\%$　(16)

4.3　实验结果

为了验证聚类分析的可靠性和有效性，将本文算法和 PSO_FCM[8]、GA_FCM[9]和 SA_FCM[10]的进行对比，对比结果如表 1 和图 4-图 7 所示。

表 1 不同算法评价结果

Table 1 Evaluation Results for Different algorithms

方法 Method	评价指标 Indexes	
	准确率 Accuracy	误判率 False
GWO_FCM	98.41%	1.59%
PSO_FCM	96.30%	3.70%
GA_FCM	96.50%	3.50%
SA_FCM	94.33%	5.67%
FCM	91.43%	8.57%

由表 1 和图 4~7 结果对比发现，本文算法的数据聚类分析准确率和误判率分别为 98.41%和 1.59%，优于 PSO_FCM、GA_FCM 和 SA_FCM 和 FCM 的数据误判率和诊断准确率，效果最优。通过本文算法和 PSO_FCM、GA_FCM 和 SA_FCM 和 FCM 的对比发现，本文算法可以有效提高数据聚类分析的准确率和降低误判率，为数据聚类分析提供新的方法和途径。

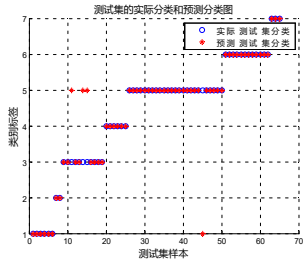

图 4 GWO_FCM 聚类结果

Fig.4 Clustering results of GWO_FCM

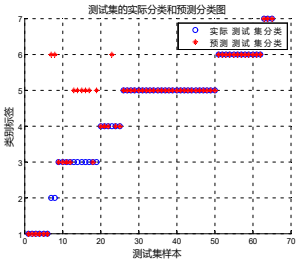

图 5 PSO_FCM 聚类结果

Fig.5 Clustering results of PSO_FCM

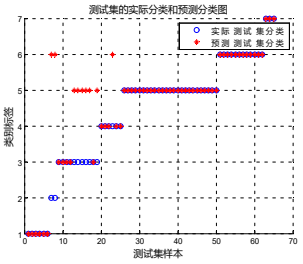

图 6 GA_FCM 聚类结果

Fig.6 Clustering results of GA_FCM

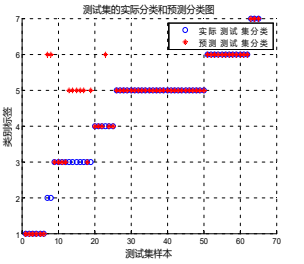

图 7 SA_FCM 聚类结果

Fig.7 Clustering results of SA_FCM

5　结 论

针对 FCM 算法聚类效果易受其初始参数影响和易陷入局部最优的缺点，将灰狼优化算法和 FCM 结合，提出一种基于 GWO 优化 FCM 的聚类分析方法。以 KDD CUP99 数据集为研究对象，研究结果表明，与 PSO、GA 和 SA 算法相比较，GWO 算法聚类分析的准确率和误判率更低，具有更快的收敛速度，效果更优，为数据聚类分析提供新的方法和途径。

参考文献

[1]　刘宜平,沈　毅.一种 FCM 聚类算法的改进与优化[J].系统工程与电子技术,2000,22(4):1-3

[2]　Mirjalili S, Mirjalili SM, Lewis A. Grey Wolf Optimizer[J]. Advances in Engineering Software, 2014,69(3):46-61

[3]　张少敏,赵　硕,王保义.基于云计算和量子粒子群算法的电力负荷曲线聚类算法研究[J].电力系统保护与控制,2014(21):93-98

[4]　Mirjalili S. How effective is the Grey Wolf optimizer in training multi-layer perceptrons[J]. Applied Intelligence, 2015,43(1):150-161

[5]　Saremi S, Mirjalili SZ, Mirjalili SM. Evolutionary population dynamics and grey wolf optimizer[J]. Neural Computing and Applications, 2015,26(5):1257-1263

[6]　陈寿文.基于 Chebyshev 映射的混沌粒子群融合 FCM 聚类算法[J].计算机应用与软件,2015,32(7):255-258

[7]　周选林,冯　晓,钟明权,等.基于遗传算法的 FCM 聚类分析在边坡稳定性评价中的应用[J].路基工程,2014(1):1-4

[8]　张永库,尹灵雪,孙劲光.基于改进的遗传算法的模糊聚类算法[J].智能系统学报,2015(4):627-635

[9]　蒋　君,徐蔚鸿,潘　楚.基于粒计算和模拟退火的 K—medoids 聚类算法[J].计算机仿真,2015,32(12):214-217

[10]　任昌荣.基于 MCQPSO-SA 优化的 KFCM 算法在入侵检测中的应用[J].计算机与现代化,2015(2):90-94