Spark 平台下聚类挖掘的智能推荐系统

钟桂凤[1],庞雄文[2],孙道宗[3],刘宇东[2]

(1.广州理工学院 计算机科学与工程学院,广东 广州 510540;2.华南师范大学 计算机学院,广东 广州 530631;
3.华南农业大学 电子工程学院,广东 广州 510642)

摘 要:为了提高智能推荐系统的性能,采用狼群优化的 K-means 聚类挖掘实现数据分类,通过协同过滤完成智能推荐。为了提高推荐效率,引入 Spark 平台多节点完成聚类和推荐。建立用户和资源的 K-means 聚类模型,采用狼群优化算法对初始类别中心点进行优化,以提高聚类准确度,根据用户和资源的类别属性获得用户-资源评分数据,最后建立协同过滤智能推荐模型。根据推荐效率要求,将推荐模型部署至 Spark 平台,实现聚类和智能推荐的分布式运算。实验证明,通过合理设置聚类中心点数目,结合 Spark 平台多节点运算,与常用推荐算法对比,所提算法可以获得更准确的推荐性能,在大规模数据的智能推荐系统中更能满足实时性要求,智能推荐效率高。

关键词:智能推荐;K-means 聚类;Spark 平台;协同过滤;狼群算法
中图分类号:TP311.13 **文章编号**:1005-9830(2021)05-0575-07
DOI:10.14177/j.cnki.32-1397n.2021.45.05.008

Intelligent recommendation system of clustering mining under Spark platform

Zhong Guifeng[1], Pang Xiongwen[2], Sun Daozong[3], Liu Yudong[2]

(1.College of Computer Science and Engineering, Guangzhou Institute of Science and Technology, Guangzhou 510540, China;
2.School of Computer, South China Normal University, Guangzhou 530631, China;
3.School of Electronic Engineering, South China Agricultural University, Guangzhou 510642, China)

Abstract: In order to improve the performance of intelligent recommendation system, the K-means clustering mining of wolf group optimization is used to achieve data classification, and intelligent recommendation is completed through collaborative filtering. In order to improve the recommendation efficiency, Spark platform multi node is introduced to complete clustering and recommendation.

收稿日期:2021-04-14 **修回日期**:2021-08-25
基金项目:国家自然科学基金(31101077);2020 年度广东省高校科研项目(2020GXJK201)
作者简介:钟桂凤(1983-),女,讲师,主要研究方向:大数据应用、数据分析与挖掘、人工智能,E-mail:109488818@qq.com。
引文格式:钟桂凤,庞雄文,孙道宗,等. Spark 平台下聚类挖掘的智能推荐系统[J]. 南京理工大学学报,2021,45(5):575-581.
投稿网址:http://zrxuebao.njust.edu.cn

Firstly, the K-means clustering model of users and resources is established, and the initial category center point is optimized by wolf group optimization algorithm to improve the clustering accuracy. The user resource scoring data is obtained according to the category attributes of users and resources, and finally, the collaborative filtering intelligent recommendation model is established. According to the recommendation efficiency, the recommendation model is deployed to Spark platform to realize the distributed operation of clustering and intelligent recommendation. The experiment shows that by setting the number of cluster centers reasonably and combining with Spark platform multi node operation, compared with the common recommendation algorithms, this algorithm can obtain more accurate recommendation performance, and can meet the real-time requirements and high efficiency of intelligent recommendation in the intelligent recommendation system with large-scale data.

Key words: intelligent recommendation; K-means clustering; Spark platform; collaborative filtering; wolf swarm algorithm

互联网高速发展,网络数据量迅速增长,用户对网络的依赖程度及数据获取便捷度的需求明显提升,用户主动搜索获取服务的方式正逐渐改变[1],平台推荐服务方式被用户青睐。对于区块链和资源共享等网络服务平台来说,数据资源需要进行聚类归档,并根据用户在平台的访问和使用习惯[2],对用户进行类别评分,然后根据相似度计算获得用户和资源的相似关系,最后为用户推荐相似度最高的资源。这种基于数据聚类的智能推荐系统是当前网络推荐系统的主流模式,这种模式因为涉及到聚类运算,在类别较多且数据规模较大的情况下,推荐的准确率和效率均会受到影响,因此多类别聚类和大规模运算效率是该推荐模式需要重点解决的问题。

当前,基于聚类挖掘的智能推荐技术研究较多,Liu 等[3]采用深度模型来实现 K-means 聚类挖掘,并用于在线学习资源推荐,能够根据用户短时间的学习习惯进行资源推荐,但是受模型限制,推荐的准确率并不高; Liu 等[4]采用深度学习算法来提升 K-means 聚类的准确率,对网站在线用户提供广告推荐,具有一定的智能推荐效果,但也出现了轻度误判; Ming 等[5]根据用户历史点歌情况,利用聚类挖掘算法实现了不同用户的歌单推荐,取得了良好效果,但其采用的方法处于封闭歌单库训练,导致适用度有一定局限。上述方法均在一定程度上提高了聚类的准确率,但是均没有效利用云计算平台来提升推荐效率。

狼群算法是近期提出的一种新型群体智能优化算法,能够解决全局寻优和局部极值问题。本文尝试引入狼群算法来对 K-means 聚类算法进行优化,以提高多类别聚类的准确率,同时引入 Spark 平台的多节点并行计算来提高聚类和推荐效率。

1 Spark 结构

Spark 作为大规模数据处理常用引擎,采用主-从节点管理模式共同完成数据处理任务,但是在功能结构上,主-从节点具有同等的运算能力,其主要结构[6,7],如图 1 所示。

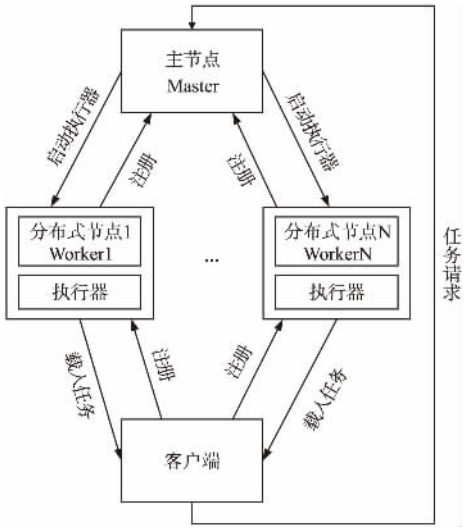

图 1 Spark 结构

除了这种主-从节点协同工作模式之外,Spark 平台还有一个优点就是大部分数据运算都在其平台节点内存的弹性分布式数据集(Resilient distributed datasets, RDD)中完成,这种方式极大地提升了数据存取效率,完成大规模数据的聚类与推荐,解决了聚类中频繁迭代造成的

运算效率不高问题,同时也解决了智能推荐的实时性问题。

2　基于狼群优化的 K-means 聚类挖掘的协同过滤推荐

2.1　K-means 算法

聚类空间中任意两点 i 和 j 的距离 S_{ij} 数学[8]表示

$$S_{ij}=\begin{cases}d_{ij} & i\neq j\\ 0 & i=j\end{cases} \tag{1}$$

设中心点 x_i 包含 n 个属性,表示方法为(x_{i1},x_{i2},…,x_{in}) 和待聚类点 x_j(x_{j1},x_{j2},…,x_{jn}) ,x_i 与 x_j 的距离为

$$d_{ij}=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2} \tag{2}$$

根据式(2) 可以计算所有待聚类的样本点至中心点距离集,根据距离 d_{ij} 来判定 x_i 与 x_j 是否同类。然后根据距离集建立聚类目标函数,求解式(3) 的最小值

$$\varepsilon=\sum_i\left\|x_i-\sum_{j,x_j\in N(x_i)}S_{ij}x_j\right\|^2 \tag{3}$$

$x_j\in N(x_i)$ 意思是:x_j 为 N 个样本点中除了中心点 x_i 的剩余样本点,满足:$\sum_{j,x_j\in N(x_i)}S_{ij}x_j=1$,$S_{ij}\geqslant 0$。

将式(3) 进一步展开得[9]

$$\varepsilon=\sum_i\left\|x_i-\sum_{j,x_j\in N(x_i)}S_{ij}x_j\right\|^2=\left\|\sum_{j,x_j\in N(x_i)}S_{ij}(x_i-x_j)\right\|^2 \tag{4}$$

最后得到目标函数为

$$\min\varepsilon$$
$$\text{s.t.}\ \sum_j S_{ij}=1,S_{ij}\geqslant 0 \tag{5}$$

K 均值聚类的效率取决于待聚类点的维度与样本数据量。一般而言,聚类的准确率和效率随着待聚类样本的数量及维度增加而降低。在处理大规模聚类精度问题和聚类时间问题仅采用 K-means算法不够,因为,除了待聚类的数据量外,K 均值聚类算法初始中心点的选择也很重要,它影响着聚类的效率,所以,有必要对 K-means 算法进行一定改进。

2.2　狼群算法

设狼群总量为 N,数据维度为 D,则第 i 只狼位置为 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{id},\cdots x_{iD})$,其中 $1\leqslant i\leqslant N$,$1\leqslant d\leqslant D$。

$$x_{id}=x_{\min}+rand*(x_{\max}-x_{\min}) \tag{6}$$

式中:$x_{\max}$ 和 $x_{\min}$ 分别表示 d 维空间的上下限。$rand$ 为 [0,1] 随机值。

根据适应度值最高的狼作为头狼,其周围的为探狼,数量为 T_{num},其游走步长[10]为

$$Step_C(d)=|\max_d-\min_d|/S \tag{7}$$

式中:$1\leqslant d\leqslant D$,S 为可设置的权重常量。

探狼位置更新

$$x_{i,d}^{C}=x_{i,d}+\sin(2\pi/h)\times Step_C(d) \tag{8}$$

式中:$i=1,2,\cdots,T_{\text{num}}$,$h$ 为游走方向数。

狼群中剩余狼移动步长

$$Step_B(d)=2\times|\max_d-\min_d|/S \tag{9}$$

位置更新为

$$x_{i,d}^{B}=x_{i,d}+Step_B(d)\cdot[s_{i,d}-x_{i,d}]/|s_{i,d}-x_{i,d}|$$

式中:$i=1,2,\cdots,N-T_{\text{num}}-1$。$s_{i,d}$ 为 d 维空间中第 i 只狼与头狼距离,$s_{i,d}\in D_d$。

$$D_d=\frac{1}{D\times\omega}\times\sum_{d=1}^{D}|\max_d-\min_d| \tag{10}$$

式中:ω 为距离因子常量。

当狼群找到猎物,头狼号令围攻,运动步长和位置更新计算方式[11]为

$$Step_W(d)=|\max_d-\min_d|/(2\times S) \tag{11}$$

$$x_{i,d}^{W}=x_{i,d}+\lambda\cdot Step_W(k)\cdot|s_d-x_{i,d}| \tag{12}$$

式中:$s_{i,d}\in D_d$,$i=1,2,\cdots,N-1$,$\lambda\in[-1,1]$ 随机值。

2.3　协同过滤推荐

通过狼群优化的 K-means 算法进行聚类后,可以获得用户对所有待推荐的资源或服务的评分,然后采用协同过滤算法进行有效推荐。

设推荐的用户集合为 $U=\{u_1,u_2,\cdots,u_m\}$,待推荐的资源集合为 $I=\{i_1,i_2,\cdots,i_n\}$,$r_{m,n}$ 表示第 m 个用户对第 n 个资源的评分,那么用户 a 和 b 的相似关系[12]为

$$sim(a,b)=\frac{\sum_{p=1}^{n}(r_{a,p}-\bar{r}_a)(r_{b,p}-\bar{r}_b)}{\sqrt{\sum_{p=1}^{n}(r_{a,p}-\bar{r}_b)^2(r_{b,p}-\bar{r}_b)^2}} \tag{13}$$

式中:$\bar{r}_a$ 和 $\bar{r}_b$ 分别表示用户 a 与 b 对所有资源的平均评分。

在协同过滤时,除了可以对用户之间的相似性进行分析之外,最重要的是需要求解用户对资源的评分,用户 j 对资源 k 的评分方法为

$$score(j,k)=\bar{r}_a+\frac{\sum_{p=1}^{n}sim(j,k)^2(r_{b,p}-\bar{r}_b)}{\sum_{p=1}^{n}sim(j,k)} \quad (14)$$

根据资源评分分数，为用户推荐评分高的资源，从而完成智能推荐。该算法的智能性体现在无需所有用户对所有资源进行评分，而是通过用户访问网络的习惯数据，采用狼群优化的K-means聚类来预测用户对资源的评分值。

2.4 聚类及推荐流程

首先，分析客户端的智能推荐任务需求，然后搭建 Spark 平台，部署合适规模的分布式节点，接着建立聚类运算模型，并通过狼群算法对初始聚类中心点进行优化，通过聚类结果获取用户-属性评分数据，最后采用协同过滤完成智能推荐。Spark 平台下聚类挖掘的智能推荐流程主要如图 2所示。

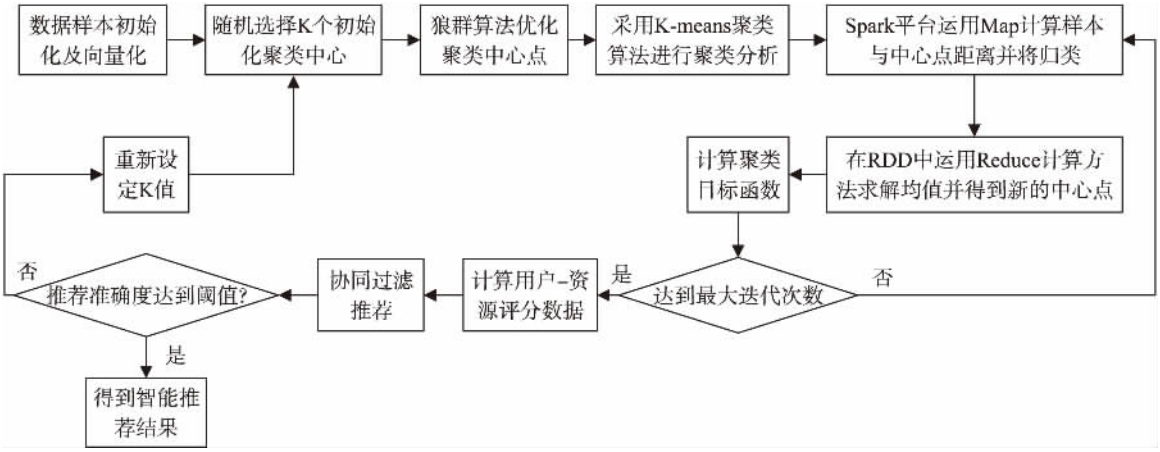

图 2 Spark 平台下智能推荐流程图

3 实例仿真

为了验证 Spark 平台下聚类挖掘的智能推荐性能，分别对公共数据集和自有数据集进行仿真，其中公共数据集为 Movie Lens 数据集，自有数据集为某在线教育平台。Movie Lens 数据集作为推荐系统仿真的经典数据集，能够很好地验证聚类挖掘的推荐性能，而在线学习平台因为用户量众多和学习资源数据量大，很容易获得大规模数据样本，充分验证 Spark 平台的推荐优势。

Spark 平台共包含 1 个 Master 节点和 9 个 Work 节点，所有节点具有相同的硬件性能。

3.1 Movie Lens 数据集仿真

为了验证狼群优化的 K-means 聚类算法在 Movie Lens 数据集的智能推荐性能，采用狼群优化的 K-means 算法完成聚类，然后通过协同过滤完成影片推荐。

表 1 Movie Lens 实验数据集

样本集	用户数据	电影数据	样本大小
Data1	1956	1683	100K
Data2	6042	3680	1MB
Data3	72001	10000	10MB

3.1.1 不同聚类中心数的推荐性能

聚类中心个数 K 对影片评分矩阵影响敏感，从而影响影片协同过滤推荐的稳定性，因此差异化设置 K，验证推荐准确率的 RMSE 值。

图 3 表明 RMSE 值随着 K 值的增加先减小后增大，当分类类别较小时，影片类别分类粒度大，因此推荐的影片与用户实际评分值偏差大，Data1 和 Data2 在 K=16 时获得了最优 RMSE，而 Data3 在 K=18 时获得了最优 RMSE 值。当继续增加 K 值后，RMSE 逐渐增大，推荐稳定性变差。因此，选择在后续针对 Movie Lens 数据集仿真时，K 取值范围设置为 [16,18]。

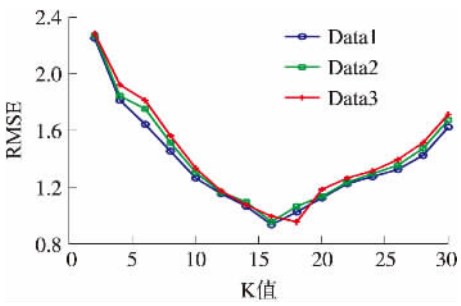

图 3 不同 K 值的推荐准确率 RMSE 值

3.1.2 基于公共数据集的推荐性能

设置 K=16，采用狼群优化的 K-means 算法

进行聚类挖掘，并采用协同过滤推荐算法对 Movie Lens 数据集训练，训练时共有 10 个节点组成 Spark 平台进行计算。

表 2　推荐性能(Movie Lens 数据集)

样本集	准确率/%			推荐时间/s
	最小值	均值	最大值	
Data1	86.913	88.539	90.062	1.156
Data2	86.752	88.613	89.435	1.197
Data3	86.913	88.784	89.391	1.213

推荐时间性能方面，容量的差别在推荐时间上表现不明显，这主要是采取了 Spark 平台的作用，对于 10 个节点来说，因 3 个样本数据容量差异导致的推荐时间变化非常小。

3.2　在线学习平台数据集仿真

为了进一步验证本文算法和 Spark 平台结合的智能推荐性能，采用在线学习平台数据集进行仿真。分别选择了粤港澳大湾区某大型在线学习平台 1 个月的用户学习数据，组建成 4 个不同容量的样本: Data1(12.95MB) ，Data2(656.82MB) ，Data3(1.42GB) ，Data4(8.03GB) 。

3.2.1　聚类结果可视化

将 4 种数据样本分别进行狼群优化 K-means 聚类。根据用户和资源的类别属性获得用户-资源评分数据，差异化设置 K 值，取推荐准确率最高 K 值作为聚类中心数; 为了直观显示聚类结果，对聚类结果进行可视化，其中 Data1 的聚类结果如图 4 所示。

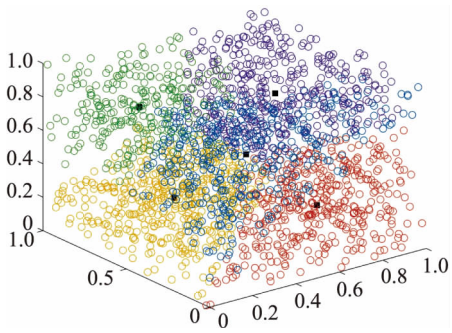

图 4　Data1 聚类结果可视化

狼群优化的 K-means 聚类算法在三维空间内将 Data1 数据集分为 5 类。根据分类结果，可以获得样本所有用户和资源属性的评分，然后再进行协同过滤计算获得推荐结果。

3.2.2　推荐性能仿真

差异化设置聚类中心数，对不同 K 值下的推荐性能进行仿真，取最优 K 值完成 5 个样本的狼群优化 K-means 和协同过滤智能推荐。

从表 3 可以看出，推荐的准确率保持在 91% 以上，准确率受样本容量的影响较小，而推荐时间随着样本容量在增加，虽然 Data3 和 Data4 样本容量量级变大，推荐时间并未有快速增长，这主要是 Spark 平台多节点运算的原因。

表 3　推荐性能(在线学习平台数据集)

样本集	样本集容量	准确率/%	推荐时间/s
Data1	12.95MB	93.495	2.817
Data2	656.82MB	92.317	7.352
Data3	1.42GB	91.464	8.663
Data4	8.03GB	91.173	12.747

3.2.3　Spark 平台的加速性能仿真

为了验证 Spark 平台对智能推荐速度的影响，求解 Spark 推荐相对于单机推荐的加速比:

$$S=\frac{T_a}{T_s} \tag{15}$$

T_a 与 T_s 为单机和 Spark 多节点的各自推荐时间。

从表 4 可以看出，当 Worker 节点数量增加，Spark 加速效果明显，样本容量越大，Worker 节点数对加速比影响越显著。Data1 的样本容量为 12.95 MB，Worker 节点达到 10 时，加密比只比单机增加了 0.001，而当样本容量为 8.03 GB 时，加速比相对于单机增加了 42.907，因此 Spark 平台提高了大容量样本的推荐效率，特别适合大规模聚类挖掘及推荐。

表 4　Spark 加速性能

样本集	样本集容量	参与计算的节点数	加速比
Data1	12.95MB	1	1.000
		5	1.001
		10	1.001
Data2	656.82MB	1	1.000
		5	1.013
		10	1.024
Data3	1.42GB	1	1.000
		5	6.473
		10	13.139
Data4	8.03GB	1	1.000
		5	22.287
		10	43.907

3.3 不同推荐算法性能仿真

为了继续验证本文算法在智能推荐系统中的性能,分别从在线学习平台的 4 个数据集中各抽取 500 个样本组建成新的数据集 Data5,将 SVM 算法[13]、深度神经网络(DNN)[14]、XGBoost 算法[15]和本文算法分别进行仿真。

如图 5 所示,4 种不同算法的推荐准确率在初期时均随着迭代次数的增加而不断提升,然后趋于稳定。推荐性能包括两个方面:准确率和收敛速度,可从这 2 个方面综合分析。首先,在推荐准确率方面,算法稳定即收敛后,本文算法和 XGBoost 算法最优,均超过了 0.9,而 SVM 算法最差,仅约为 0.7;其次,在收敛速度方面,SVM 表现最优为 190 次,本文算法次之,约为 230 次,XGBoost 算法最差。根据 4 种算法的综合推荐性能对比来看,本文算法在准确率和收敛速度 2 个方面均排名靠前,这说明其对在线学习样本的综合推荐性能最佳。

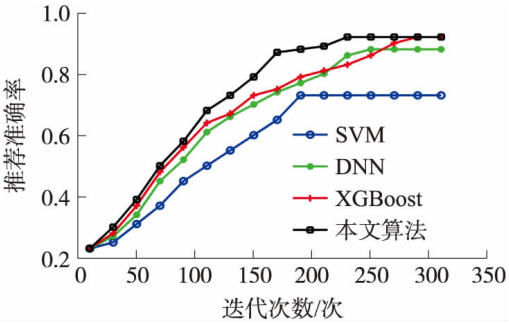

图 5 不同算法的推荐性能

4 结束语

本文采用狼群优化的 K-means 算法完成聚类挖掘,并采用协同过滤算法完成智能推荐,推荐准确率高;为了解决大规模数据的推荐问题,引入 Spark 平台多节点共同完成聚类和推荐,提高了智能推荐效率。后续研究将进一步优化聚类参数及 Spark 节点的自适应加入,以提高智能推荐准确率,同时节省节点计算资源。

参考文献:

[1] Helmers C, Krishnan P, Patnam M. Attention and saliency on the Internet: Evidence from an online recommendation system [J]. Journal of Economic Behavior & Organization, 2019, 161: 216-242.

[2] Orús C, Gurrea R, Ibáñez-Sánchez S. The impact of consumers' positive online recommendations on the omnichannel webrooming experience [J]. Spanish Journal of Marketing-ESIC, 2019, 23(3): 397-414.

[3] Liu Jie, Zhang Hong, Liu Zihui. Research on online learning resource recommendation method based on wide & deep and elmo model [J]. Journal of Physics: Conference Series, 2020, 1437: 012015.

[4] Liu D R, Liao Yushan, Chung Y H, et al. Advertisement recommendation based on personal interests and ad push fairness [J]. Kybernetes, 2019, 48(8): 1586-1605.

[5] He Ming, Guo Hao, Lv G, et al. Leveraging proficiency and preference for online Karaoke recommendation [J]. Frontiers of Computer Science, 2020, 14(2): 273-290.

[6] Tang Bing, Kang Linyao, Zhang Li, et al. Collaborative filtering recommendation using nonnegative matrix factorization in GPU-accelerated spark platform [J]. Scientific Programming, 2021, 2021: 1-15.

[7] Yang Yan, Yu Juan, Yang Mengfan, et al. Probabilistic modeling of renewable energy source based on Spark platform with large-scale sample data [J]. International Transactions on Electrical Energy Systems, 2019, 29(3): e2759.

[8] 王法玉,刘志强. Spark 框架下分布式 K-means 算法优化方法 [J]. 计算机工程与设计, 2019, 40(6): 1595-1600.
Wang Fayu, Liu Zhiqiang. Optimization method of distributed K-means algorithm based on Spark [J]. Computer Engineering and Design, 2019, 40(6): 1595-1600.

[9] 黄松,邱建林. 改进的遗传 K-means 算法及其应用 [J]. 计算机工程与设计, 2020, 41(6): 1617-1623.
Huang Song, Qiu Jianlin. Improvement of K-means clustering algorithm based on genetic algorithm and its application [J]. Computer Engineering and Design, 2020, 41(6): 1617-1623.

[10] Subramanian P, Sahayaraj J M, Senthilkumar S, et al. A hybrid grey wolf and crow search optimization algorithm-based optimal cluster head selection scheme for wireless sensor networks [J]. Wireless Personal Communications, 2020, 113(2): 905-925.

[11] Al-Din B N, Manasrah A M, Noaman S A. A novel approach by using a new algorithm: Wolf algorithm as a new technique in cryptography [J]. Webology, 2020,

17(2) : 817-826.

[12] 王源龙,孙卫真,向勇. 基于 Spark 的混合协同过滤算法改进与实现[J]. 计算机应用研究, 2019, 36(3) : 855-860.
Wang Yuanlong, Sun Weizhen, Xiang Yong. New improvement and implementation of hybrid collaborative filtering algorithm based on Spark platform [J]. Application Research of Computers, 2019,36(3) : 855-860.

[13] 崔建双,车梦然. 基于多分类支持向量机的优化算法智能推荐系统与实证分析[J]. 计算机工程与科学,2019,41(1) : 153-160.
Cui Jianshuang, Che Mengran. An intelligent recommendation system for optimization algorithms based on multi-classification support vector machine and its empirical analysis [J]. Computer Engineering & Science,2019,41(1) : 153-160.

[14] 田德红,何建敏. 基于变异粒子群优化与深度神经网络的航空弹药消耗预测模型[J]. 南京理工大学学报,2018,42(6) : 716-721,726.
Tian Dehong, He Jianmin. Aviation ammunition consumption prediction model based on mutated particle swarm optimization and deep neural network [J]. Journal of Nanjing University of Science and Technology,2018,42(6) : 716-721,726.

[15] Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree [J]. Information Sciences, 2021,572: 522-542.