计算机科学
COMPUTER SCIENCE

一种适于多分类问题的支持向量机加速方法

陈景年

引用本文

陈景年. 一种适于多分类问题的支持向量机加速方法[J]. 计算机科学, 2022, 49(6A): 297-300.

CHEN Jing-nian. Acceleration of SVM for Multi-class Classification[J]. Computer Science, 2022, 49(6A): 297-300.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210400149

一种适于多分类问题的支持向量机加速方法

陈景年
山东财经大学信息与计算科学系　济南 250014

摘　要　支持向量机因具有卓越的分类效果和坚实的理论基础而成为了近年来模式识别、机器学习以及数据挖掘等领域中最重要的分类方法之一。然而,其训练时间会随样本增多而明显增长,并且在处理多分类问题时模型训练会更加复杂。为解决上述问题,给出了一种适于多分类问题的训练数据快速约简方法 MOIS。该方法以聚类中心为参照点,在删除掉冗余训练样本的同时,选择起决定作用的边界样本来大幅度约简训练数据,并消减类别间的分布不均衡问题。实验结果表明,MOIS 在保持甚至提高支持向量机分类效果的同时,能大幅提高训练效率。例如,在 Optdigit 数据集上,利用所提方法使分类准确率由 98.94%提高到 99.05%的同时,训练时间缩短到原来的 15%;又如,在 HCL2000 前 100 类构成的数据集上,在准确率略有提高的情况下(由 99.29%提高到 99.30%),训练时间更是大幅缩短到不足原来的 6%。另外,MOIS 本身具有很高的运行效率。

关键词:支持向量机;多分类;数据约简;聚类;样本选择
中图法分类号　TP391

Acceleration of SVM for Multi-class Classification

CHEN Jing-nian
Department of Information and Computing Science, Shandong University of Finance and Economics, Jinan 250014, China

Abstract　With excellent classification effect and solid theoretical foundation, support vector machines have become one of the most important classification method in the field of pattern recognition, machine learning and data mining in recent years. However, their training time becomes much longer with the increase of training instances. In the case of multi-class classification, the training process will become even more complex. To deal with above problems, a fast data reduction method named as MOIS is proposed for multi-class classification. With cluster centers being used as reference points, redundent instances can be deleted, bound instances crucial for the trainning can be selected, and the distribution imbalance between classes can also be relieved by the proposed method. Experiments show that MOIS can enormously improve the training efficiency while keeping or even improving the classification accuracy. For example, on Optdigit dataset, the classification accuracy is increased from 98.94% to 99.05%, while the training time is reduced to 0.15% of the original. What's more, on the dataset formed by the first 100 classes of HCL2000, the training time of the proposed method is reduced to less than 6% of original, while the accuracy is improved slightly from 99.29% to 99.30%. Furthermore, MIOS is highly efficient.

Keywords　Support vector machines, Multi-class classification, Data reduction, Clustering, Instance seletion

1　引言

在模式识别、机器学习以及数据挖掘等领域,支持向量机(Support Vector Machine,SVM)[1]因具有卓越的分类效果和坚实的理论基础而成为近年来备受瞩目的分类方法。结构风险最小化和凸二次规划的使用,使 SVM 较其他分类方法有着更强的分类性能。核方法的使用,使 SVM 对非线性可分问题也具有很好的分类效果。SVM 已被成功应用到文字识别[2]、图像分类[3]、金融预测[4]、医疗诊断[5]等许多科技领域。

然而,正是训练过程中二次规划的应用,使得 SVM 训练复杂度随训练样本数的增加而显著增加。一般情况下,对于含有 n 个样本的训练集,在其上面进行 SVM 训练的时间复杂度为 $O(n^3)$。另外,SVM 是针对二分类问题提出的,对于多分类问题,通常的做法是将多分类问题通过一对一或一对多的方式转化为多个二分类问题来解决。一对一的转换方式需要在任两个类之间训练一个 SVM 分类模型,且分类过程较为复杂。一对多方式需要训练的模型要少得多,且分类过程较简单,但每个模型都是在严重不均衡的数据集上训练而得的。不管采取哪种转化方式,模型训练都会更加复杂。

近年来,随着信息技术的发展,数据集的规模也在不断增大,其中多分类问题广泛存在,高训练复杂度已成为 SVM 在许多实际应用中的主要瓶颈。如何提高 SVM 的训练效率,尤其对多分类问题而言,是一个重要且亟待解决的研究课题。

为降低 SVM 的训练复杂度,研究人员主要从两种途径进行探索,即提高二次规划效率和样本选择。为提高二次规划的效率,研究人员试图将整个训练集上的二次规划过程分解为一系列的小规模的优化过程。这方面比较典型的算法有 SMO(Sequential Minimal Optimization)[6],SOR(Successive

基金项目:国家自然科学基金(61773325)
This work was supported by the National Natural Science Foundation of China(61773325).
通信作者:陈景年(jnchen06@163.com)

Over Relaxation)[7], Chunking[8], LIBSVM (LIBrary for SVM)[9]等算法。基于样本选择的方法则是从训练集中选择对训练结果起决定作用的样本来构成规模较小的训练子集，并在其上训练 SVM 分类模型。由于 SVM 的训练复杂度高度依赖于训练样本数，因此样本选择的方法对提高训练效率具有更加显著的作用。正因如此，本文将遵循这一思路，以聚类中心为参照点，通过样本选择构造一种适合多分类问题的 SVM 加速算法。本文首先对通过样本选择来提高 SVM 效率的相关研究进行概述；然后介绍所提算法，通过实验验证所提算法的有效性；最后对工作进行总结，并提出下一步的研究设想。

2 相关工作

在 SVM 的训练过程中，只有作为支持向量的边界样本才对训练结果起作用[10]。其余样本，要么是不起作用的冗余样本，要么是对训练结果起损害作用的噪声样本。基于样本选择的 SVM 加速方法就是要删除掉冗余和噪声样本，选择可能是支持向量的边界样本。基于这一思路，Almeida 等[11]提出了 SVM-KM 算法。该算法首先对训练样本进行聚类。一个簇中如果只包含同一类样本，则其中的所有样本用该簇的聚类中心来替代；如果一个簇中包含不同类的样本，则该簇中的所有样本都被保留。SVM-KM 具有较高的效率，但许多情况下选择效果不够理想。Li 等[12]提出了一种训练样本删除算法，首先用一个小规模的样本子集训练得到一个初始的 SVM，然后将原样本集中离初始 SVM 分类超平面较远的训练样本删除。该方法可以删除掉一些对分类无关紧要的样本，但选择效果依赖于初始的 SVM。Shin 等[13]提出了一种快速的样本选择方法 NPPS(Neighborhood Property Based Pattern Selection Algorithm)，利用训练样本的邻域信息进行分析，NPPS 选择位于分类超平面附近的样本构成最终的训练集。该方法可明显提高训练速度，但选择结果很容易受到噪声数据的干扰。Angiulli 等利用一致子集(Consistent Subset)给出了样本选择算法 FCNN(Fast Condensed Nearest Neighbor Rule)[14]。该算法精简样本的幅度较大，容易引起分类精度的降低。Li 等[15]利用样本的 K 近邻类的类分布信息和几何特征给出了 BEPS(Border Edge Pattern Selection)算法，以选择关键样本。该算法可充分利用训练集中样本的分布信息，但容易选择较多的样本，且需要给定 4 个超参数的值，给算法应用增加了难度。Kim 等[16]从训练集中随机选择一定比例的样本构成多个训练子集，并在每个子集上训练一个 SVM 模型，然后利用这些模型对每个样本进行评价，并根据评价结果选择样本。由于每个样本的评价结果是由随机选择的样本子集上训练的 SVM 产生的，因此选择结果难免带有一定的随机性。

上述研究在提高 SVM 对二分类问题的效率方面取得了明显成效。然而，对于多分类问题，这些算法在效率和效果方面往往难以取得满意的效果。针对这一问题，本文构建了一种适于多分类问题的快速样本选择方法 MOIS(Multi-classification Oriented Instance Selection)。对于多分类问题，该算法在效率和效果方面都较已有算法有明显优势。

3 MOIS 算法

3.1 MOIS 算法框架

如前文所述，为了将 SVM 用于多分类问题，需把原问题转化为多个二分类问题。相比一对一转化方式，一对多方式不仅训练模型要少得多，而且分类过程也更简单高效，只需解决此方式所引起的正负类间的不均衡问题即可。因此，MOIS 算法将采用一对多的转化方式。

假定训练集 A 是由 L 类样本所组成的，A 中样本总数为 N，第 $c(1\leqslant c\leqslant L)$类样本的数目为 N_c。按照一对多转化方式，A 中的每一类样本都轮流做一次正类样本，其余样本共同作为当下的负类样本。下文考虑将第 $l(1\leqslant l\leqslant L)$类作为当前正类时的边界样本选择过程，其中 S_l 表示选择的结果。

首先，利用某种聚类方法(本文采用 k 均值聚类方法)对当前的正类样本——第 l 类样本进行聚类，并假设 k 个聚类中心为 $C_1, C_2, \cdots, C_k$。然后，对于每个 $C_i(1\leqslant i\leqslant k)$，计算其到每个样本 x 的距离 $d(x, C_i)$。最后，从第 l 类中选择 $d(x, C_i)$较大的一定比例的正类样本。同时，从其他每个类中选择 $d(x, C_i)$较小的一定比例的负类样本。这一做法的依据是：对于正例而言，距离正类的簇中心越近，越可能为内点；否则越可能为边界点。对负例而言，距离正类簇中心越近，是边界点的可能性就越大。

3.2 MOIS 中参数的确定

接下来的关键问题是如何确定总的样本选择比例以及正例和负例的选择比例。

3.2.1 总样本选择比例及聚类簇数的确定

本节给出了一种简单有效的方法来确定总的样本选择比例以及聚类簇数。该方法允许用户根据计算资源和数据集的大小来选取合适的比例和簇数。一般来说，对于样本较少的小数据集，其稀疏程度较高，选取的样本比例也较高。相反，对于样本数较大而维数不高的数据集，因样本较为稠密，往往选取较低的比例即可取得理想的效果。因此，我们可以先根据数据集的大小确定比例值的范围，然后通过实验验证在该范围内选取一个合适的值。比如，对于较小的数据集，可在{0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7}中通过实验验证选取一个恰当的值。

对于正类中的簇数 k，也是根据正类的大小来选取。当正类样本较少时，k 的值一般也取得小一些；反之，则需要较大的 k 值。通过实验发现，一般 k 在 1 到 7 之间取值时，即可得到理想的选择效果。

3.2.2 正例和负例选择比例的确定

在确定正例和负例的选择比例时，考虑到当前负类是由第 l 类之外的其他多个类所组成，极易出现正例数显著少于负例数的情况，设置的样本选择比例应有助于消除正负类间的这种高度不均衡性。同时，考虑到由多个类所组成的负类的边界一般更复杂，选择的负例数还应适当多于正例数。比如 MOIS 算法中的做法，令负例数为正例数的两倍。

另外，在选择负例时，应从整个负类中选择，还是应从构成负类的每个类中分别选择呢？由于组成负类的多个类往往大小不一，因此应从其中每个类中选择负例，且选择的负例数应与该类的大小成正比。假设选择的总负例数为 N^s_-，则从第 $i(1\leqslant i\leqslant L, i\neq l)$类中选择的负例数 N^s_i 为：

$$N^s_i=\frac{N^s_-\cdot N_i}{\sum\limits_{1\leqslant c\leqslant L, c\neq l} N_c} \tag{1}$$

综合上述分析，MOIS 算法的步骤如算法 1 所示。

算法 1 MOIS 算法

输入:由分成 L 类的共 N 个样本组成的训练集 T,样本选择比例 r

输出:第 l(1≤l≤L)类作为当前正类时的选择结果 S_l

步骤 1 对作为当前正类的第 l(1≤l≤L)类样本进行聚类,得到 k 个聚类中心 $C_1,C_2,\cdots,C_k$

步骤 2 对于每个聚类中心 C_i(1≤i≤k),计算其到每个样本 x 的距离 $d(x,C_i)$。

步骤 3 计算要选择的正例数 N^s_+,N^s_- 如下:

$$N^s_+=\min\{\frac{1}{3}N\cdot r,N_l\},N^s_-=\frac{2}{3}N\cdot r。$$

步骤 4 从当前正类中选择 N^s_+ 个 $d(x,C_i)$较大的样本放入 S_l。

步骤 5 从第 i(1≤i≤L,i≠l)类中选择 N^s_i 个 $d(x,C_i)$较小的负例放入 S_l。其中,N^s_i 由式(1)确定。

在得到的选择结果 S_l 上训练 SVM,即可得到第 l 个分类模型,令 l 取 1,2,…,L,便会得到 L 个分类模型。

3.3 MOIS 算法的复杂度分析

算法 1 中的步骤 1 中,在正类上的聚类过程的复杂度为 $O(TkN/L)$。其中,k 的值不大于 7,为聚类中心的数目;T 为聚类过程中的循环次数。当 k 设为 1 时,复杂度归为 $O(N/L)$。算法 1 中的步骤 2 的复杂度为 $O(kN)$。算法 1 中步骤 3—步骤 5 主要是对每类样本按照距离度量进行排序,复杂度为 $O(kN\log(N/L))$。综上分析,MOIS 的复杂度为 $O(N\log(N/L))$。由此可见,MOIS 具有很高的效率。

下文将通过实验验证 MOIS 算法的有效性。

4 实验验证

为验证 MOIS 算法的有效性,实验中将其与最具代表性的 3 种算法 NPPS[13],FCNN[14] 以及 BEPS[15] 进行比较。这 3 种算法都被用来对 SVM 进行加速,并取得了显著的效果。实验中将从下列几个方面对上述 4 个参与比较的算法进行评估:1)对分类效果的影响;2)对数据集的精简幅度;3)对训练效率的提升效果;4)算法本身的执行效率。

4.1 实验数据集

实验中采用了多个标准的多分类数据集和一个实际应用中的手写汉字识别数据集。表 1 列出了这些数据集的样本数(Size)、特征数(#Fea)、类别数(#Cls)、训练样本数(#Trn)以及测试样本数(#Tes)等信息。除了 HCL2000 外,其余数据集均来自 UCI 机器学习数据库[17],并且每个数据集已被分成了训练集和测试集。HCL2000 是一个由 3755 个类构成的中文手写字符集[18],每个类包含 1000 个样本,700 个用于训练,300 个用于测试。由于在整个 HCL2000 上对 SVM 的训练和测试时间太长,这里只取其前 100 类样本用于实验。通过提取 8-方向梯度特征[19],每个样本的特征数为 512。

表 1 实验数据集

Table 1 Details of experimental data sets

Data set	Size	#Fea.	#Cls	#Trn.	#Tes.
Dermatology	358	34	6	282	76
Glass	214	9	6	168	46
HCL2000	100000	512	100	70000	30000
Iris	150	4	3	117	33
Isolet	7797	617	26	6238	1559
Letter	20000	16	26	16000	4000
Optdigits	5620	64	10	3823	1797
Pendigit	10992	16	10	7494	3498
USPS	9298	256	10	7291	2007

4.2 实验参数设置

为避免在高维数据集上运行时间过长,对 HCL2000,Isolet 以及 USPS 等高维数据集采取了惯常采用的降维措施。其中,HCL2000 通过 LDA 方法降到 99 维,Isolet 和 USPS 利用 PCA 分别降到 150 维和 80 维。

实验中采用的 SVM 的核函数一律为式(2)定义的高斯函数。

$$k(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2) \quad (2)$$

在每个数据集上的误差限参数以及式(2)中的参数 σ 都通过实验验证做了优化设置。MCIS 中的选择比例与聚类数按 3.2 节中的说明进行设置。NPPS 算法中的近邻数、BEPS 中的 k_b,k_e,λ 以及 γ 等参数都按原文所说的方法进行设置。这里,只有 FCNN 算法无需参数设置。

4.3 实验结果

为比较各种算法在实验数据集上的性能,表 2—表 5 分别列出了利用各种算法在每个数据集上得到的分类准确率、样本选择比例、模型训练时间以及选择样本的时间。表中的"All"表示利用整个训练集进行训练的情形。

表 2 在实验数据集上的分类准确率

Table 2 Classification accuracy on experimental data sets

(单位:%)

数据集	算法				
	All	BEPS	FCNN	MOIS	NPPS
Dermatology	100.00	88.16	92.11	**98.68**	94.74
Glass	69.57	56.52	56.52	**65.22**	58.70
HCL2000	99.29	99.17	99.02	**99.30**	98.28
Iris	100.00	100.00	90.91	**100.00**	63.64
Isolet	96.34	95.32	95.77	**96.06**	94.74
Letter	97.97	93.65	95.67	**97.40**	86.42
Optdigits	98.94	98.83	97.61	**99.05**	97.83
Pendigit	98.80	98.26	96.80	**98.46**	98.34
USPS	95.81	95.17	95.52	94.67	**95.62**

表 3 在实验数据集上的样本选择比例

Table 3 Ratio of selected instances on experimental data sets

数据集	算法			
	BEPS	FCNN	MOIS	NPPS
Dermatology	0.64	**0.33**	0.44	0.55
Glass	**0.39**	0.45	0.52	0.63
HCL2000	0.988	**0.070**	**0.070**	0.170
Iris	0.31	**0.15**	0.43	0.32
Isolet	0.91	**0.30**	0.40	0.55
Letter	0.46	**0.18**	0.35	0.47
Optdigits	0.93	**0.09**	0.25	0.41
Pendigit	0.38	**0.05**	0.40	0.29
USPS	0.96	**0.12**	0.34	0.47

表 4 在实验数据集上的训练时间

Table 4 Training time on experimental data sets

(单位:s)

数据集	算法				
	All	BEPS	FCNN	MOIS	NPPS
Dermatology	0.20	0.16	0.11	**0.08**	0.14
Glass	0.17	0.05	**0.03**	0.06	0.08
HCL2000	25173.9	23043.6	7616.8	**1456.2**	3378.6
Iris	0.03	0.03	0.03	**0.01**	**0.01**
Isolet	290.97	231.56	**76.8**	89.99	116.86
Letter	2139.5	885.29	**228.06**	596.42	713.03
Optdigits	65.03	59.13	**4.23**	9.75	25.22
Pendigit	36.22	12.88	**2.06**	12.73	12.36
USPS	197.78	142.23	**14.53**	43.42	78.49

表 5 4 种选择算法在数据集上的运行时间

Table 5 Runtime of four algorithms on experimental data sets

(单位:s)

数据集	算法			
	BEPS	FCNN	MOIS	NPPS
HCL2000	969.59	6869.90	**60.98**	1087.50
Isolet	7.76	22.19	**1.97**	5.88
Letter	16.78	54.28	**0.87**	18.42
Optdigits	4.53	2.86	**0.17**	6.50
Pendigit	3.92	1.47	**0.16**	5.05
USPS	8.55	4.55	**0.59**	11.52

由表 2 可以发现,在 4 种样本选择算法中,MOIS 在除 USPS 外的所有实验数据集上都保持了最高的分类准确率,更甚至在 HCL2000 和 Optdigits 数据集上 MOIS 使分类准确率较在原数据集上还有所提高。只是在 USPS 上,MOIS 比其他 3 种算法得到的分类准确率略低。反观另外 3 种算法,它们都没有很好地保持分类准确率。

由表 3 列出的样本选择比例来看,总体上 MOIS 的选择比例仅高于 FCNN,而明显低于 NPPS 和 BEPS。值得注意的是,FCNN 的高精简度容易导致分类准确率的明显下降。

观察表 4 列出的 SVM 训练时间可以发现,4 种样本选择算法都能明显缩短训练时间。相比之下,MOIS 对训练时间的缩短能力仅略次于 FCNN,而明显优于 NPPS 和 BEPS。

由于 4 种选择算法在小规模数据集上的运行时间非常短暂,因此不容易察觉它们在运行时间上的明显差别,表 5 只列出了这些算法在较大规模数据集上的运行时间。通过表 5 可以发现,运行时间上 MOIS 显著短于其他算法,一般只有其余算法的几分之一,甚至几十分之一,这充分表明 MOIS 的运行效率显著高于其他算法。

结束语 作为模式识别、机器学习以及数据挖掘等领域中最重要的分类方法之一,支持向量机具有卓越的分类效果。然而,该方法的模型训练时间会随样本增多而明显增长,尤其在处理多分类问题时模型训练会更加复杂,不仅训练的模型显著增多,模型的训练效果也有待提高。为解决上述问题,本文给出了一种适于多分类问题的训练数据快速约简算法 MOIS。该方法首先对当前正类进行聚类,然后以得到的聚类中心为参照点,在删除掉冗余样本的同时选出起决定作用的边界样本,并且通过适当控制正负例的选择比例来消减类别间的分布不均衡现象。与以往基于聚类的数据约简方法不同,MOIS 由于只对一类样本进行聚类,聚类速度大大超过了以往算法。另外,删除冗余样本和选择边界样本并举,使得 MOIS 的数据约简效果更优。与以往性能卓越的几种算法进行实验比较发现,MOIS 在保持支持向量机分类效果方面明显优于其他算法,在运行效率上也显著更高。

参 考 文 献

[1] VAPNIK V. The nature of statistical learning theory[M]. New York:Springer,1995.

[2] DONG J,KRZYZAK A,SUEN C Y. Fast SVM training algorithm with decomposition on very large data sets [J]. IEEE Trans. Pattern Analysis and Machine Intelligence,2005,27(4):603-618.

[3] YANG B Q,GUAN X P,ZHU J W,et al. SVMs multi-class loss feedback based discriminative dictionary learning for image classification[J]. Pattern Recognition,2020,112(12):76-90.

[4] ZHANG X D,LI A,PAN R. Stock trend prediction based on new status box method and adaboost probabilistic support vector machine [J]. Applied Soft Computing,2016,49:385-398.

[5] RAMÍREZ J,GÓRRIZ J,SALAS-GONZALEZ D,et al. Computer-aided diagnosis of alzheimer's type dementia combining support vector machines and discriminant set of features [J]. Information Sciences,2013,237:59-72.

[6] KEERTHI S S,SHEVADE S K,BHATTACHARYYA C,et al. Improvements to platt's SMO algorithm for SVM classifier design [J]. Neural Computation,2001,13(3):637-649.

[7] MANGASARIAN O L,MUSICANT D R. Successive overrelaxation for support vector machines [J]. IEEE Transactions on Neural Networks,1999,10(5):1032-1037.

[8] VAPNIK V. Estimation of dependences based on empirical data[M]. New York:Springer,2006.

[9] CHANG C C,LIN C J. LIBSVM:A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(3):1-27.

[10] BURGES C J. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery,1998,2:121-167.

[11] ALMEIDA M B,BRAGA A P,BRAGA J P. SVM-KM:Speeding SVMs learning with a priori cluster selection and k-means [C]// Brazilian symposium on neural networks. Brazil Computer Society,2000:162-167.

[12] LI H L,WANG C H,YUAN B Z,et al. A Learning Strategy of SVM Used to Large Training Set [J]. Chinese Journal of Computers,2004,27(5):715-719.

[13] SHIN H,CHO S. Neighborhood property based pattern selection for support vector machines[J]. Neural Computation,2007,19(3):816-855.

[14] ANGIULLI F,ASTORINO A. Scaling up support vector machines using nearest neighbor condensation[J]. IEEE Transactions on Neural Networks,2010,21(2):351-357.

[15] LI Y,MAGUIRE L. Selecting critical patterns based on local geometrical and statistical information [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2011,33(6):1189-1201.

[16] KIM D,KANG S,CHO S. Expected margin-based pattern selection for support vector machines[J]. Expert Systems With Applications,2020,139:1-12.

[17] HETTICH S,BLAKE C L,MERZ C J. UCI Repository of machine learning databases[EB/OL]. http//www. ics. uci. edu/~mlearn/ MLRepository. html.

[18] ZHANG H,GUO J,CHEN G,et al. HCL2000—A Large-scale Handwritten Chinese Character Database for Handwritten Character Recognition[C]// International Conference on Document Analysis and Recognition. IEEE Computer Society,2009:286-289.

[19] LIU C L,NAKASHIMA K,SAKO H,et al. Handwritten digit recognition:investigation of normalization and feature extraction techniques[J]. Pattern Recognition,2004,37(2):265-279.

CHEN Jing-nian, born in 1970, Ph. D, professor, supervisor, is a senior member of China Computer Federation. His main research interests include big data analysis, intelligent information processing.