广西科学院学报
Journal of Guangxi Academy of Sciences
2017,33(1):32～39
Vol. 33, No. 1 February 2017

网络优先数字出版时间:2017-02-23 【DOI】10.13657/j.cnki.gxkxyxb.20170223.001
网络优先数字出版地址:http://www.cnki.net/kcms/detail/45.1075.N.20170223.0855.002.html

基于 PSO 的 WFCM 算法研究及其在医保欺诈行为发现中的应用*

Study on WFCM Algorithm based on PSO and Its Application in Identifying Medicare Fraud

李 华,陈宁江
LI Hua, CHEN Ningjiang

(广西大学计算机与电子信息学院,广西南宁 530004)
(School of Computer, Electronics and Information in Guangxi University, Nanning, Guangxi, 530004, China)

摘要:【目的】在没有先验知识的前提下,采用基于粒子群优化算法(PSO)的加权模糊 C-均值(WFCM)聚类算法,从 30 多万条记录的医疗保险数据中挖掘出疑似医疗保险欺诈的记录。**【方法】**首先,引用改进的欧式距离、相似性函数以及交叉熵函数并通过 PSO 算法极小化交叉熵函数,对属性权重进行分析;其次,选取 Calinski-Harabasz(CH)有效性指标,展开聚类有效性的研究;然后,基于数据预处理的结果将数据运用于 PSO 算法,不断更新得到各属性的权重,并运用聚类有效性评价中的 CH 有效性指标来动态估计最佳聚类个数,提高 FCM 聚类的速度;最后,将属性权重和最佳聚类数应用于 FCM 聚类算法,根据隶属度矩阵聚类得到疑似医疗保险欺诈结果。**【结果】**基于上述研究方法,本研究根据最后的隶属度矩阵来进行聚类分析。**【结论】**将优化的权重应用于加权 FCM 聚类算法与聚类有效性评价,既提高了聚类算法的高效性,又避免了主观评价对分类的影响。

关键词:PSO WFCM CH 有效性指标 医保欺诈
中图分类号:TP311.13 **文献标识码:**A **文章编号:**1002-7378(2017)01-0032-08

Abstract:【Objective】This paper aims to find the records of suspected medicare fraud from over 30 million records by using the Weighted Fuzzy C-Means clustering algorithm based on particle swarm optimization (PSO) algorithm with the absence of prior knowledge. **【Methods】**Firstly, the improved Euclidean Distance, similarity function and cross entropy function are introduced and the entropy function is minimized by PSO algorithm to analyze the attribute weight. Secondly, the validity index of CH (Calinski-Harabasz) is selected, and the study of validity of clustering is carried out. Thirdly, the data is applied to the PSO algorithm based on the results of data preprocessing, constantly updated to get the weight of each attribute, and the optimal numbers of clusters are estimated dynamically by validity index of CH, in order to increase the speed of FCM. Finally, the attribute weights and the optimal clustering numbers are applied to the FCM clustering algorithm, and the results of suspected medical insurance fraud are obtained according to the membership matrix. **【Results】**Based on the above method, the final membership matrix is used for carrying out cluster analysis. **【Conclusion】**This

收稿日期:2016-11-26
修回日期:2016-12-07
作者简介:李 华(1994－),男,硕士,主要从事云计算与大数据、网络软件工程研究,E-mail:820402018@qq.com。
*国家自然科学基金项目(61363003)资助。

paper shows the running efficiency of clustering algorithms can be improved, and the influence of subjective evaluation for classification can be avoided by applying the weights to the WFCM clustering algorithm and clustering validity.

Key words: PSO, WFCM, validity index of CH, medicare fraud

0　引言

【研究意义】医保欺诈成为当前医保领域的一个难题。医保数据具有保险类数据的特征,数据量大、数据属性类型多、动态性强,且包括众多的数据以及庞大而复杂的属性关系。大数据背景下,运用相关聚类算法主动挖掘医疗保险欺诈行为,显得尤其重要。**【前人研究进展】**当前研究大多采用神经网络等有监督的学习方法或者未赋予指标权重,主观性较强。比如王熙照等[1]根据模糊等价关系将给定对象划分成一些等价类,在一定程度上会导致划分不精确等问题。**【本研究切入点】**在经典的聚类分析算法中,无论是采用欧氏距离还是马氏距离,都未将所有指标赋予各自的权重。本研究就医保欺诈行为的主动发现这一问题进行探讨,并采用非辅助学习(无监督学习)方法。通过比较分析,采用聚类分析算法,对数据进行分类处理,认为出现的孤立点为疑似欺诈点。**【拟解决的关键问题】**本研究将先构造交叉熵函数,再通过粒子群优化算法不断更新属性权重,得到最终满足交叉熵函数 CFuzziness(ω) 的条件。另外,为提高 FCM 聚类算法的速度,将通过聚类有效性评价来估计聚类个数,最后将属性权重以及聚类个数应用于经典 FCM 聚类算法进行聚类分析。

1　基于 PSO 的加权 FCM 算法

传统聚类方法忽略各因素的权重差异,因此本研究引入交叉熵函数 CFuzziness(ω),该函数刻画随着权重 ω 的改变,分类模糊程度的变化。通过调整 ω 的值,使分类的模糊程度尽量小。ω 的最优解使同一类的点尽量靠近,不同类的点尽量远离,即 ω 应对应于 CFuzziness(ω) 取最小值时的最优解。常用的梯度下降算法[1]的时间复杂度大,且容易陷入局部最优解,因此本研究采用粒子群优化算法[2]求解 ω 的最优解。另外,运用聚类有效性评价[3]中的 CH 有效性指标[4]来确定最佳聚类数,然后再根据聚类数来进行下一步的 FCM 聚类算法[5]。

1.1　符号定义

相关的符号含义如表 1 所示。

1.2　改进的欧氏距离

引入基于属性权重的欧氏距离 $d_{pq}^{(\omega)}$,称之为改进的欧氏距离,其定义为

$$d_{pq}^{(\omega)}=\sqrt{\sum_{k=1}^{m}\omega_k(X_{pk}-X_{qk})^2}\,, \tag{1}$$

其中 $\omega=\{\omega_1,\omega_2,\cdots,\omega_m\}$ 是权重向量,对应于属性向量;$\omega_k\in[0,1]$ 为第 k 维属性对应的权重,在相似性度量中,第 k 维属性的作用与 ω_k 的成正比,则类间距离可定义为

$$d(s,t)=\frac{1}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}d_{X_{si}X_{tj}}^{(\omega)}\,。 \tag{2}$$

表 1　符号说明表

Table 1　Symbol description

符号 Symbol	符号解释 Description	符号 Symbol	符号解释 Description
X	医保记录数据集 Medicare record data set	ω_l	每一维的属性权重 The weight of each dimension
X_{ij}	X_i 的第 j 维属性的数据 The data of the j-th dimensional attribute of X_i	C_i	X 的簇(或子类) The cluster of X (or subclass)
d_{pq}	通常使用的欧式距离 The usual Euclidean Distance	c	数据集 X 的中心点 The center point of the data set X
$d_{pq}^{(\omega)}$	基于属性权重的欧氏距离 Euclidean Distance based on attribute weights	c_i	簇 C_i 的中心点 The center point of cluster C_i
S_{pq}	欧氏距离下的相似性关系 The similarity under Euclidean Distance	μ_{ij}	第 j 个数据从属第 i 类隶属度 The j-th data is subordinate to the i-th class membership
$S_{pq}^{(\omega)}$	加权欧氏距离下的相似性 Similarity under weighted Euclidean Distance	$v_i^{(b+1)}$	第 b 次迭代后第 i 个簇的中心 The center of the i-th cluster after the b-th iteration

1.3 粒子群优化(PSO)算法

1.3.1 交叉熵函数

为得到属性权重 ω_k，引入交叉熵函数 CFuzziness(ω)[1]，采用改进的欧氏距离 $d_{pq}^{(\omega)}$ 作为度量指标后，在相似性关系不变(即如果 $S_{pq}>0.5$，则 $S_{pq}^{(\omega)}>0.5$；如果 $S_{pq}<0.5$，则 $S_{pq}^{(\omega)}<0.5$) 的前提下，相似性度量定义也相应变化，表示为

$$S_{pq}^{(\omega)}=\frac{1}{1+\beta\times d_{pq}^{(\omega)}}。\tag{3}$$

常数 $\beta\in[0,1]$，β 值通过调整，使得 $S_{pq}^{(\omega)}$ 能够近似于正态分布散落在[0,1] 内。由下式得到 β 的近似值[1]：

$$\frac{2}{n(n-1)}\sum_{q<p}S_{pq}=\frac{2}{n(n-1)}\sum_{q<p}\frac{1}{1+\beta\times d_{pq}}=0.5。\tag{4}$$

为让聚类结果具有模糊性相对较小的性质，通过调整数据集各属性权重，使得相似数据间的距离更小，而不相似数据间的距离更大，即通过综合函数来评价各个点之间在各属性作用下的相似程度，使总体范围内的模糊性最小。基于以上分析，引进交叉熵函数[1]：

$$\text{CFuzziness}(\omega)=\frac{-2}{n(n-1)}\sum_{q<p}\frac{1}{2}(S_{pq}^{(\omega)}\times\log S_{pq}+(1-S_{pq}^{(\omega)})\times\log(1-S_{pq}))。\tag{5}$$

1.3.2 PSO 算法计算属性权重

本研究利用粒子群优化算法[6]来极小化交叉熵 CFuzziness(ω)。假设目标搜索空间是一个 D 维数据集，PSO 算法初始化为一个群落，这个群落由 N 个粒子组成，其中第 i 个粒子用一个 D 维的向量 $X_i=\{x_{i1},x_{i2},\cdots,x_{iD}\}$，$i=1,2,\cdots,N$ 来表示，也用一个 D 维的向量描述第 i 个粒子“飞行”的速度，记作：

$$V_i=\{v_{i1},v_{i2},\cdots,v_{iD}\},i=1,2,\cdots,N,\tag{6}$$

迄今为止第 i 个粒子搜索到的最佳位置称为个体极值，记作：

$$P_{best}=\{p_{i1},p_{i2},\cdots,p_{iD}\},i=1,2,\cdots,N,\tag{7}$$

迄今为止整个粒子群搜索到的最佳位置为全局极值，记作：

$$g_{best}=\{p_{g1},p_{g2},\cdots,p_{gD}\},g=1,2,\cdots,D。\tag{8}$$

粒子在飞行时不断地根据公式(9) 和(10) 来决策，并通过下面两个公式[7] 来更新位置 x_{id} 和速度 V_{id}：

$$V_{id}=w\times v_{id}+c_1r_1(p_{id}-x_{id})+c_2r_2(p_{gd}-x_{id}),\tag{9}$$

$$x_{id}=x_{id}+v_{id},i=1,2,\cdots,N,g=1,2,\cdots,D。\tag{10}$$

粒子群优化算法优化属性权重 ω 值的流程[8]如下：

1)初始化粒子群中的群体规模 N，各粒子速度 V_i 和粒子位置 X_i；

2) 通过 $\sum V_iX_i$ 计算得到各粒子的适应度值，即目标函数值 $F_{it}[i]$；

3) 比较各粒子的适应度值 $F_{it}[i]$ 与个体极值 $P_{best}(i)$，若 $F_{it}[i]>P_{best}(i)$，则 $F_{it}[i]$ 取代 $P_{best}(i)$；

4) 比较各粒子的适应度值 $F_{it}[i]$ 与全局极值 g_{best}，若 $F_{it}[i]>g_{best}(i)$，则用 $F_{it}[i]$ 取代 g_{best}；

5) 根据公式(9) 和(10) 更新粒子的速度 V_i 和位置 X_i；

6)满足最大循环次数或最小误差时退出循环，否则返回步骤 2)。

1.4 聚类有效性评价

通常，在聚类之前需要划分聚类数，而对于数据集划分的类数往往是未知的，也不容易得到。聚类有效性评价就是通过选定聚类有效性指标[9]来评价聚类算法和聚类个数的优劣，使得聚类有效性指标最优时的聚类个数为最佳聚类个数。本研究选取 Calinski-Harabasz(CH)有效性指标[3]来进行评价。

对于 m 维数据集 $X=\{X_1,X_2,\cdots,X_n\}$，n 为数据集中数据对象个数，其中 $X_i=\{x^1,x^2,\cdots,x^m\}$ 表示第 i 个数据对象。划分式聚类算法[10]将数据集 X 分为 NC 个子集 $X=\{C_1,C_2,\cdots,C_{NC}\}$，$C_i$ 为 X 的簇(子类)，用 c 表示数据集 X 的中心点，c_i 表示簇 C_i 的中心点，n_i 表示簇 C_i 中的数据对象的个数，$d(X_i,X_j)$ 表示数据对象 i 和数据对象 j 之间的距离。

于是 CH 有效性指标的定义如下[11]：

$$\text{CH}(NC)=\frac{\frac{1}{NC-1}\sum_{i=1}^{NC}n_id^2(c_i,c)}{\frac{1}{n-NC}\sum_{i=1}^{NC}\sum_{x\in c_i}d^2(x,c_i)}。\tag{11}$$

采用估计聚类个数工具箱(NCT)[11]进行聚类有效性评价，NCT 包括 4 个外部有效性指标和 8 种内部有效性指标。主文件设计为如何使用聚类算法、有效性指标和方法来估计聚类个数。

1.5 加权模糊 C-均值(WFCM)

K-Means[12]隶属度取值为 0 或者 1，根据“类内误差的平方和最小”这个准则不断迭代。设 $X=\{X_1,X_2,\cdots,X_n\}\subset R^P$，$R^P$ 表示 p 维的向量空间，NC 表示聚类数，隶属度 μ_{ij} 表示第 j 个数据样本从属于第 i 类的程度，目标函数为

$$(\min)J_1=\sum_{i=1}^{NC}\sum_{j=1}^{N}\mu_{ij}\parallel x_j-v_i\parallel^2。\qquad(12)$$

K-Means 算法简单快速,时间复杂度为 $O(nK\tau)$,呈线性关系,因此对于大数据特别是数值型大数据的处理比较有效,但是 K-means 也有以下 3 个方面的缺点:

①对于初始聚类中心有很强的敏感性以及依赖性;

② K 的选取需要一定的先验知识;

③不适用于数据分布呈凸形状的聚类,对孤立点很敏感。

典型 K-Means 聚类算的基本根据是"类内加权误差平方和最小化"[13],其目标函数为

$$(\min)J_2=\sum_{i=1}^{NC}\sum_{k=1}^{N}\mu_{ik}^2\parallel x_k-v_i\parallel^2。\qquad(13)$$

Bezdek 对于上一目标函数推广到更为普遍的形式:

$$(\min)J_m=\sum_{i=1}^{NC}\sum_{k=1}^{N}\mu_{ik}^m\parallel x_k-v_i\parallel^2。\qquad(14)$$

建立在 FCM 算法的基础上,李伯年[14]将 FCM 算法里的欧式距离进行优化,推广到广义的欧式距离,即加权模糊 C-均值聚类算法(WFCM),适用于大数据的排序与聚类。WFCM 算法的基本步骤类似于 FCM 算法步骤,只是在目标函数上存在区别以及更新隶属度矩阵要考虑属性的权重。

WFCM 的目标函数为

$$(\min)J_m=\sum_{i=1}^{NC}\sum_{k=1}^{N}\mu_{ik}^m\omega_l\parallel x_k-v_i\parallel^2,l=1,\cdots,m,(m\text{ 为数据的维数})。\qquad(15)$$

WFCM 算法的具体步骤如下:

1)声明聚类的簇的个数 NC,取值范围为[2,n],n 为数据个数,给定停止迭代的阈值 ε 以及初始的聚类中心 V^0、迭代累计数 $b=0$,另外数据维数 m 的取值为 2;

2)初始隶属度矩阵 $U^0=[u_{ij}(0)]$ 根据先验知识确定,并用下列公式计算及更新隶属度矩阵 $U^b=[u_{ij}]$:

$$u_{ij}^{(b+1)}=\Big[\sum_{k=1}^{NC}\Big(\frac{(d_{ij}^{(\omega)})^2}{(d_{kj}^{(\omega)})^2}\Big)^{\frac{1}{m-1}}\Big]^{-1},1\leqslant i\leqslant NC,1\leqslant j\leqslant n。\qquad(16)$$

其中 $d_{ij}^{(\omega)}=\sqrt{\sum_{l=1}^{m}\omega_l(x_{jl}-v_{il})^2}$,$d_{kj}^{(\omega)}=\sqrt{\sum_{l=1}^{m}\omega_l(x_{jl}-v_{kl})^2}$;

3)用下列公式修正聚类中心 V^{b+1}:

$$v_i^{(b+1)}=\frac{\sum_{j=1}^{n}(u_{ij}^{(b)})m_{x_j}}{\sum_{j=1}^{n}(u_{ij}^{(b)})m},1\leqslant i\leqslant NC;\qquad(17)$$

4)如果 $\parallel V^b-V^{b+1}\parallel\leqslant\varepsilon$,停止算法并输出隶属度矩阵 U 以及聚类中心 V,否则令 $b=b+1$,并回到步骤 2)继续执行。停止迭代以后,若 $\max(u_{ij})=u_{kj}$,则 $x_j\in$ 第 k 类。

2　医保欺诈行为主动发现中的应用

2.1　数据分析及预处理

医保数据是深圳医院一个月 30 多万条记录的数据,具有庞大而复杂的属性关系,且含有众多缺失或属性不清晰的数据。因此,先进行数据预处理。定性定量地从病人、医生、科室的角度进行分析:从病人的角度,包括一张卡一定时间内反复多次拿药、单张账单费用较高、一人持多卡即一卡多用这 3 种情况。再结合疑似欺诈账单号,从医生的角度和科室的角度来进行分析,重点是从病人的角度来分析。

经过数据准备阶段以后,得到两个新表 PAPMI_IDNAME3 和 Bills。首先,用 Delete_noNAME3 去处理表 PAPMI_IDNAME3,生成不含有医保号为 1 的表 PAPMI_IDNAME3_new。其次,根据 PAPMI_IDNAME3_new 删除 Bills 中病人 ID 不存在于 PAPMI_IDNAME3_new 的记录,用 Delete_BillsnoNAME3 去实现。再次,因为账单 Bills_new 中同一个账单号可以有多条记录,所以应将同一个账单号的多条记录叠加一起。观察数据,知道表中除第 3 列(总价)的数据不同,其余都相同,依次对第 3 列的数据相加后为总费用,仍存储于第 3 列,使用函数 Count_price 处理后生成数据表 Bills_new2。最后,运用函数 Count_frequency 统计每个病人 ID 购药的次数,生成表 Bills_new3,其新增的第 6 列为统计出来的每个病人 ID 购药的次数。

为方便直观地了解本阶段的处理过程,用 Visio 绘制数据流模型图如图 1 所示。

2.2　粒子群优化(PSO)算法计算属性权重

经过数据预处理阶段后,生成包括 59 081 条数据记录的表 Bills_new3,其 6 列数据分别是执行科室、病人 ID、总费用、下医嘱医生 ID、账单号、购药次数。为了采用粒子群优化算法求解权重 ω 的最优解,主函数 Calcu_weight 调用 6 个 M 文件,分别为 PSOstep.m、CFBeta.m、Gwank.m、ISwarm.m、

PProcess. m、standard. m。

主函数 Calcu_weight 为

```
>>bills=xlsread('Bills_new3. xls');
>>DATA=bills(1:10000,:);
>>DATA=standard(DATA);
>>DIS=pdist(DATA);
>>DDIS=squareform(DIS);
>>Scope=[0,1;0,1;0,1;0,1;0,1;0,1];
>>w=PProcess(30,5,Scope,@ISwarm,@PSOstep,@Gwank,0,0,1000);
```

基于以上程序，得到各项属性的权重如表 2 所示。

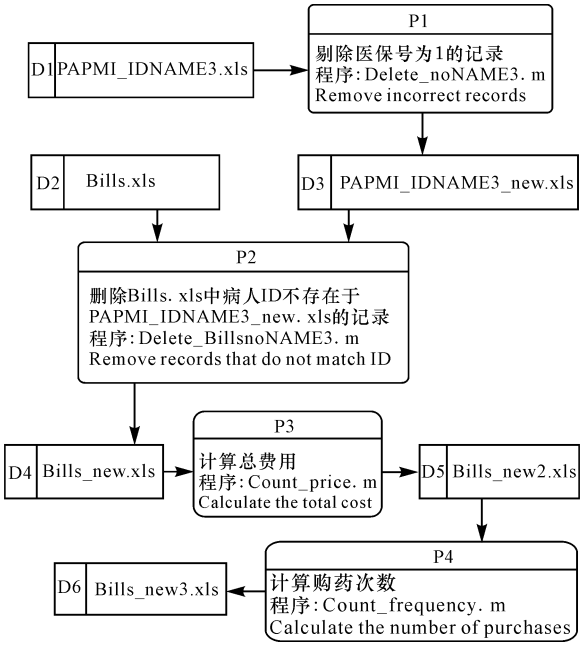

图 1 数据预处理阶段数据流模型

Fig. 1 Data flow model of the data preprocessing stage

2.3 聚类有效性评价

因为估计聚类个数工具箱(NCT)的使用中，利用 zeros (n,n) 创建高维数据，但是由于内存溢出的限制，必须对 Bills_new3 拆分后运算，然后依次对它们运行 NCT 估计聚类个数。

表 3 全部 data 的估计聚类个数结果

Table 3 The estimated number of clustering results for all data

	data1	data2	data3	data4	data5	data6	data7	data8	data9	data10
CH	66843	80631	113325	78190	79635	83233	70934	80883	80361	66727
c	10	11	11	15	14	16	9	11	10	13

2.4 加权模糊 C-均值聚类(WFCM)

基于数据表 Bills_new3、属性权重 ω 以及最佳聚类个数 c，对数据表运用加权模糊 C-均值聚类方法进行聚类，编写 4 个 M 文件：WFCMClust. m、distwfcm. m、stepwfcm. m、initwfcm. m，其中主函数 WFCMClust. m 分别调用其它两个函数。4 个 M 文件的作用分别如下：

①WFCMClust. m：采用加权模糊 C-均值对数

表 2 各属性权重表

Table 2 Attribute weight

指标 Index	权重 Weights
执行科室 Departments	0.100 7
病人 ID Patient ID	0.015 3
总费用 Total cost	0.496 6
下医嘱医生 ID Doctor ID	0.136 5
账单号 Account number	0.143 0
购药次数 Times	0.415 4

估计聚类个数工具箱(NCT)[13]包括 4 个外部有效性指标和 8 种内部有效性指标，其中包括 Calinski-Harabasz(CH)有效性指标[4]，本研究仅利用 NCT 中的 CH 指标。通过 MATLAB 对 data1 使用 NCT 估计聚类个数的部分输出结果如图 2 所示。

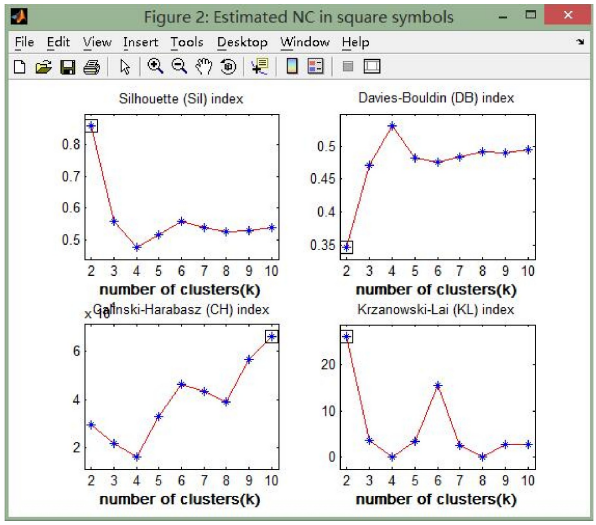

图 2 MATLAB 使用 NCT 估计聚类个数输出结果

Fig. 2 The NCT estimated number of cluster outputs by MATLAB

观察可知，CH 有效性指标显示当聚类个数为 10 时，CH 值最大，即类的自身紧密，类与类之间分散，即更优的聚类结果。同理，再依次对 data2，…，data10 的数据进行聚类有效性评价，c 为最佳聚类数。结果(省略小数)整理如表 3 所示。

据集 data 聚为 cluster_n 类;

②distwfcm. m:循环计算得所有数据点与每一个聚类中心的距离;

③stepwfcm. m:迭代计算出新的隶属度矩阵、目标函数值以及新的聚类中心;

④initwfcm. m:初始化隶属度函数矩阵。

将结果隶属度矩阵 U 提取出来并整理成表 WFCM_data1. xls,部分数据如表 4 所示。其中,表的行代表数据记录,列代表聚类后的簇编号,每一行中最大的目标函数值所在的列即为该数据行所归属的类。

表 4　WFCM 算法输出的隶属度矩阵

Table 4　The membership matrix of WFCM algorithm output

0.0385	0.2226	0.0832	0.1584	0.0203	0.0305	0.0396	0.0392	0.3310	0.0367
0.0639	0.3397	0.1668	0.0746	0.0250	0.0348	0.0330	0.0382	0.1724	0.0516
0.2102	0.1366	0.3551	0.0321	0.0343	0.0371	0.0230	0.0308	0.0587	0.0821
0.3168	0.1019	0.2917	0.0276	0.0378	0.0372	0.0214	0.0291	0.0476	0.0889
0.0246	0.0935	0.0438	0.4538	0.0165	0.0253	0.0457	0.0367	0.2345	0.0257
…	…	…	…	…	…	…	…	…	…

2.5　聚类分析

对 WFCM_data1 进行聚类分析,目的是根据隶属度矩阵进行聚类划分,划分的原则是每一行中最大的目标函数值所在的列即为该数据行所归属的类,代码如下:

```
>>data=xlsread('WFCM_data1');
>>[max_data,index]=max(data,[],2);
>>plot(index,'*');  %绘制散点图分布,目的是大体上观察聚类效果
>>Bills_new3=xlsread('Bills_new3');  %录入原始数据
>>Bills_new3=[Bills_new3,index];  %将簇编号添加到最后一列,即第 7 列
>>sortrows(Bills_new3,7);%将表 Bills_new3 按第 7 列排序,方便选出疑似欺诈记录
>>xlswrite('Result',Bills_new3);
```

运行以上程序后,输出结果为 Result. xls,其中第 7 列已经排序。

3　实例分析

根据之前的假设,医保欺诈行为不存在模仿与学习,呈孤立点状分布,即不涉及医保欺诈的账单记录归在一类中,而其他涉嫌医保欺诈的账单记录则归在其他类里。类似于以上 data1 的处理过程,对 data2～data10 的数据用 WFCM 算法以及聚类分析后,整理成 169 条记录的 result. xls,部分记录见表 5。

本研究的数据是 2014 年 1 月 1 日至 2014 年 1 月 31 日深圳市全部医院的医疗保险数据,年龄这一属性应该列入定量的标准中。利用病人资料的身份证号码变成计算年龄,绘出各个年龄阶段所占比例的情况(图 3)。

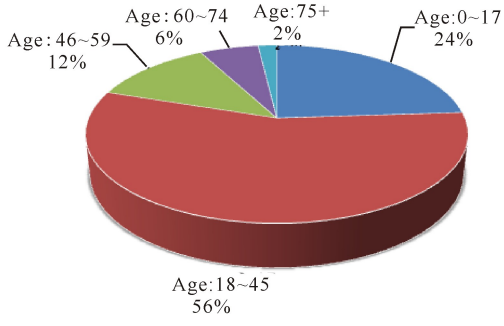

图 3　各年龄阶段所占比例

Fig. 3　The pie chart of the age

表 5　根据隶属度矩阵进行聚类分析的输出结果

Table 5　The result of clustering analysis according to membership matrix

序号 No.	账单号 Account number	病人 ID Patient ID	执行科室 Departments	下医嘱医生 ID Doctor ID	总费用(元) Total cost(¥)	购药次数 Times
1	5041943	644604	2	142	1 170.6	1
2	5042603	163696	191	1180	64.52	25
3	5042604	163696	191	1180	16	25
4	5061954	649842	191	1029	1 373.45	1
5	5062350	650134	191	1155	3 645.2	1
…	…	…	…	…	…	…

从一定程度上考虑，即使有监护人，年龄太小或者年龄太老的病人涉嫌医保欺诈的可能性都很小，因此编写 M 函数 deal_age. m 为 169 条记录的 result. xls 添加年龄，并删除 0～17 岁及 75 岁以上病人的记录，最终输出 127 条嫌疑记录，部分结果见表 6。

表 6 聚类分析最终结果——疑似医保欺诈记录

Table 6 Clustering analysis results-suspected medical insurance fraud records

序号 No.	账单号 Account Number	病人 ID Patient ID	执行科室 Departments	下医嘱医生 ID Doctor ID	总费用(元) Total cost(¥)	购药次数 Times	年龄 Age
1	5407195	36911	2	3312	43.28	2	28
2	5399943	92297	2	3312	15.68	1	30
3	5408696	92494	2	126	1 163.77	7	41
4	5294411	159988	502	1028	271.75	10	66
5	5294574	159988	502	1028	11	10	66
…	…	…	…	…	…	…	…

医生和科室也可能是医保欺诈问题的对象，因此应从科室的角度和医生的角度来发现医保欺诈记录。涉嫌医保欺诈科室通常为下医嘱科室，因此本研究重点分析下医嘱科室信息，考虑到某些科室需要长期接待同一病人或者属于高频次使用的科室，所以还需利用疑似账单号筛选出与之相关的科室，并统计出与这些疑似医保欺诈的账单号关联的次数来确定嫌疑科室。用 delete_unsuspected. m 删除账单号不存在于疑似欺诈账单号 Suspected 中的相关记录，然后重新通过数据透视图对所有科室绘制柱形图，如图 4 所示。

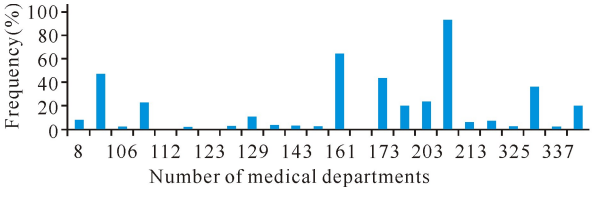

图 4 关联可疑账单号后的下医嘱科室交易频次

Fig. 4 The frequency column chart of medical advice department associated with the suspicious bill number

从图 4 中可以看出，某些科室的交易频次相对差别存在较大差异，频次较高的科室可作为重点排查和整治的目标，这才能准确地划分出的嫌疑科室。

同样利用 delete_unsuspected. m 处理的结果，通过数据透视图对所有科室绘制柱形图，频次较高的医生可作为重点排查和整治的目标，这才能准确地划分出嫌疑医生，并以此作为划分嫌疑医生的标准(图 5)。

4 结论

研究表明，将 PSO 算法优化的权重应用于 WFCM 聚类算法与聚类有效性评价，用聚类有效性

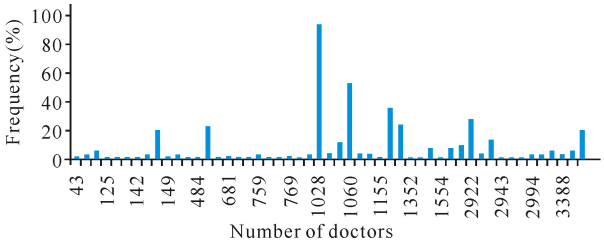

图 5 关联可疑账单号后的医生交易频次

Fig. 5 The frequency column chart of doctor associated with the suspicious bill number

评价估计最佳聚类数代替人工对聚类数的经验取值，既提高了聚类算法的高效性，又避免了主观评价对分类的影响。将研究结果应用于 30 万条医保欺诈数据的主动发现，得出疑似医保欺诈的 127 条记录。然而在研究的过程由于技术和篇幅的有限，也由于本文所提供的数据有限性和属性粗糙性，还有大量的未完成的探索性及尝试性的工作没有体现在论文中，但是对于此课题的探索和研究依然没有终止。

参考文献：

[1] 王熙照，王丽娟，王利伟. 传递闭包聚类中的模糊性分析[J]. 计算机工程与应用，2003，39(18)：92-94，129.
WANG X Z，WANG L J，WANG L W. The fuzziness analysis of transitive closure clustering[J]. Computer Engineering and Applications，2003，39(18)：92-94，129.

[2] 李宁. 粒子群优化算法的理论分析与应用研究[D]. 武汉：华中科技大学，2006.
LI N. Analysis and application of particle swarm optimization[D]. Wuhan：Huazhong University of Science and Technology，2006.

[3] BEZDEK J C. Cluster validity with fuzzy sets[J]. Jour-

nal of Cybernetics,1974,3(3):58-73.

[4] CALIŃSKI T, HARABASZ J. A dendrite method for cluster analysis [J]. Communications in Statistics, 1974,3(1):1-27.

[5] 李明华,刘全,刘忠,等.数据挖掘中聚类算法的新发展[J].计算机应用研究,2008,25(1):13-17.
LI M H, LIU Q, LIU Z, et al. New developments of clustering methods in data mining[J]. Application Research of Computers,2008,25(1):13-17.

[6] 王纵虎,刘志镜,陈东辉.基于粒子群优化的模糊 C-均值聚类算法研究[J].计算机科学,2012,39(9):166-169.
WANG Z H,LIU Z J,CHEN D H. Research of PSO-based fuzzy C-means clustering algorithm[J]. Computer Science,2012,39(9):166-169.

[7] 武妍,徐敏.一种改进的粒子群优化算法[J].计算机工程与应用,2006,42(33):40-42.
WU Y, XU M. Modified particle swarm optimization algorithm[J]. Computer Engineering and Applications, 2006,42(33):40-42.

[8] 刘逸.粒子群优化算法的改进及应用研究[D].西安:西安电子科技大学,2012.
LIU Y. Improvements and applications of particle swarm optimization algorithm[D]. Xi'an: Xidian University,2012.

[9] 刘燕驰,高学东,国宏伟,等.聚类有效性的组合评价方法[J].计算机工程与应用,2011,47(19):15-17.
LIU Y C, GAO X D, GUO H W, et al. Ensembling clustering validation indices[J]. Computer Engineering and Applications,2011,47(19):15-17.

[10] 孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,2008,19(1):48-61.
SUN J G, LIU J, ZHAO L Y. Clustering algorithms research[J]. Journal of Software,2008,19(1):48-61.

[11] WANG K J. (Simple) Tool for estimating the number of clusters[EB/OL]. 2016-04-10, http://www.mathworks.com/matlabcentral/fileexchange/13916.

[12] SULAIMAN S N, ISA N A M. Adaptive fuzzy-K-means clustering algorithm for image segmentation [J]. IEEE Transaction on Consumer Electronics, 2010,56(14):2661-2668.

[13] DUNN J C. Well-separated clusters and the optimal fuzzy partitions [J]. Journal of Cybernetics, 1974, 4(1):95-104.

[14] 李柏年.加权模糊 C-均值聚类[J].模糊系统与数学,2007,21(1):106-110.
LI B N. Weigh on cluster fuzzy C-mean[J]. Fuzzy Systems and Mathematics,2007,21(1):106-110.

(责任编辑:米慧芝)