Computer Science and Application 计算机科学与应用, 2017, 7(1), 88-99
Published Online January 2017 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2017.71011

Hans汉斯

Detecting Abnormal Use of Drugs Based on Coupled Relationships

Lizhen Wang[1,2], Junli Lu[2], Shikun Deng[1], Jing Zhang[2]

[1]School of Science and Engineering, Dianchi College of Yunnan University, Kunming Yunnan
[2]School of Information Science and Engineering, Yunnan University, Kunming Yunnan
Email: lzhwang@ynu.edu.cn

Received: Jan. 6th, 2017; accepted: Jan. 23rd, 2017; published: Jan. 26th, 2017

Abstract

In recent years, the doctor-patient relationship has received extensive attention. How to exactly mine the abnormal use of drugs is the vital to constrain doctors and relieve the doctor-patient relationship. This paper presents a general framework for detecting the abnormal use of drugs. The framework integrates coupled similarity and Chameleon clustering algorithm, and consists of three stages, qualitative analysis, quantitative analysis and abnormal detection. We conduct extensive experiments on real-world prescription data. The experiments evaluate that the framework can efficiently detect abnormal use of drugs.

Keywords

Coupled Relationships, Clustering, Abnormal Analysis

基于耦合关系的医生用药异常分析

王丽珍[1,2]，芦俊丽[2]，邓世昆[1]，张　静[2]

[1]云南大学滇池学院，理工学院，云南　昆明
[2]云南大学，信息学院，云南　昆明
Email: lzhwang@ynu.edu.cn

收稿日期：2017年1月6日；录用日期：2017年1月23日；发布日期：2017年1月26日

摘　要

近些年来，医患关系受到广泛关注。如何准确地挖掘异常用药是制约医生和缓减医患关系的关键。本文提出了一种检测医生用药异常的总体框架。该框架集成处方数据的耦合相似度度量和变色龙聚类算法，

文章引用: 王丽珍, 芦俊丽, 邓世昆, 张静. 基于耦合关系的医生用药异常分析[J]. 计算机科学与应用, 2017, 7(1): 88-99. http://dx.doi.org/10.12677/csa.2017.71011

并分为三个阶段，定性分析，定量分析和异常检测。在真实处方数据上进行了充分的实验，实验验证了该框架能够有效地检测出医生用药异常。

关键词

耦合关系，聚类，异常分析

1. 引言

近些年来，医患关系越来越引起人们的重视，尤其是对医生滥开检查单、过度用药、开大处方等问题。究其原因，一是医学的复杂性和医生的诊治水平因素；二是经济原因。医院过于市场化的发展，医务人员的收入与医院的经济效益直接关联，开单提成、药物回扣等以药养医行为屡见不鲜；三是法律规章制度。例如在医疗事故鉴定法中规定了医生在鉴定过程中举症倒置的制度，这可能导致医生对病人做大量检查、开具较多药物等“防御性”的行为[1]。

针对上述问题，本文提出一种检测医生用药异常的框架。此框架包含三个阶段，第一阶段，定性分析，将处方划分为相似团体，在相似团体下进行医生用药异常检测。第二阶段，定量分析，首先，针对复杂的处方数据，提出了基于耦合关系的处方相似度度量方法。然后，应用变色龙算法对处方数据进行聚类。第三阶段，异常检测，应用自定义的聚类指标检测医生用药异常。

大多数相似性度量都是基于数据之间相互独立的假设，但是在处方数据中，药物之间存在紧密配合与相互影响的关系，即耦合关系。考虑了耦合关系之后的医生处方用药异常检测，能够更准确地挖掘到医生用药异常情况。

论文其余部分的组织如下：第 2 节描述相关工作和相关概念，包括耦合相似度计算和变色龙聚类等。第 3 节给出了基于耦合关系的医生用药异常分析框架，包括定性分析，定量分析和异常检测。第 4 节在真实处方数据上进行了充分的实验，验证了所提出的框架和挖掘方法的有效性。

2. 相关工作和相关概念

2.1. 相关工作

相似度用于衡量对象之间的相似程度，是数据挖掘和自然语言处理的基础，在诸如聚类、异常点检测等数据挖掘技术中广泛使用。由于描述数据对象的属性多种多样，相似度计算方式也分为很多种。计算数值属性的相似度常用欧几里得距离、曼哈顿距离和闵可夫斯基距离。离散属性有二值属性和多值属性，Jaccard 系数[2]是计算具有非对称二元属性的对象间的相似度的常用方法。序数属性是指其属性值之间具有意义的序或排位，而相继值之间的量值是未知的[3]。余弦相似度是查询检索中常用的相似度计算方法，可用来计算文档之间的相似度，也可以计算词条间的相似度[4] [5]。耦合相似度是不承认属性间以及对象间的相互独立，认为属性间以及对象间存在耦合关系。耦合相似度可以对离散属性[6]和数值属性[7]进行度量，针对不同类型的数据，使用的具体度量公式不同。关于离散属性的耦合相似度的具体内容在 2.2.1 节有详细介绍。

聚类分析简称聚类，根据数据对象及其关系的信息，将数据对象划分成子集的过程[8]。每个子集是一个簇，簇中的数据对象更加相似。因为没有提供类标号信息，所以聚类分析有时被称为无监督分类。通常情况下，主要的基本聚类算法可以分为层次聚类算法、划分式聚类算法、基于密度和网格的聚类算法和其他聚类算法等。变色龙算法[9]通过一个图划分方法将数据对象聚类为大量相对较小的子簇，然后用层次聚类算法通过反复合并子簇得到结果簇。此算法利用动态模型的层次聚类方法，不需要提前设定参数，可发现任意形状和密度的簇。K-means 聚类算法[10]是最经典的划分式聚类算法的代表，但此算法通常会在获得一个局部最优值时终止，仅适合对数值型数据聚类，只适用于聚类结果为凸形(即类簇为凸形)的数据集。为克服 K-means 聚类算法的不足，研究者们提出了 K-modes-CGC 算法[11]，K-modes-Huang 算法[12]，K-means-CP 算法[13]等很多新的改进的 K-means 算法。DBSCAN 算法[14]是一个比较有代表性的基于密度的聚类算法，与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。Derya 等人[15]对 DBSCAN 进行了与辨识核对象、噪音对象和邻近类簇相关的 3 个边缘扩展，进而提出一种新的基于密度的聚类算法。

聚类分析常被用于异常检测。目前异常检测已经应用到诈骗检测、入侵检测、医疗、纳税等方面[16]。在股市中，一群隐藏的操纵者之间相互合作操纵证券价格，CHMM-CBA 方法用耦合关系来检测操纵行为。此外，一种更通用的 CBA 框架通过捕获更广泛的耦合关系去检测基于团体的市场操纵行为[17]。在医疗方面，针对社会上存在的一些不法分子在医疗行为上虚构事实，隐瞒真相，以骗取医保基金或医疗待遇的情况，潘[18]、王[19]和梁等[20]建立数据挖掘模型，预测各种医疗消费费用，并通过一系列的数据分析和挖掘，检测到医疗消费费用的异常记录。但是，将聚类分析应用到医生用药的异常检测的相关研究还很少。本文引入耦合相似度度量和变色龙聚类算法，检测医生用药异常。

2.2. 相关概念

本节主要介绍在定量分析中采用的耦合相似度[6]计算和变色龙聚类算法[9]。

2.2.1 耦合相似度

本小节以信息表为例，介绍如何计算属性值的耦合相似度(CASV)，以及对象的耦合相似度(CASO) [6]。信息表如表 1 所示。$u_i\left(1\le i\le 6\right)$ 表示第 i 个对象，$a_j\left(1\le j\le 3\right)$ 表示第 j 个属性。

使用耦合关系进行相似度度量时，不承认信息表中属性间、对象间相互独立的假定。耦合分为属性内耦合和属性间耦合。

1) 属性内耦合相似度(**IaASV**)

同一个属性内，属性值出现的次数差异反映的是值的频率分布，同一个属性内的两个值出现的频率越相近，它们的相似度越高[3]。两个对象 u_x 和 u_y 的第 j 属性的属性值 v_j^x 和 v_j^y 的属性内耦合相似度(**IaASV**)可以通过下式计算：

$$\delta_j^{Ia}\left(v_j^x,v_j^y\right)=\frac{\left|G_j\left(\left\{v_j^x\right\}\right)\right|\cdot\left|G_j\left(\left\{v_j^y\right\}\right)\right|}{\left|G_j\left(\left\{v_j^x\right\}\right)\right|+\left|G_j\left(\left\{v_j^y\right\}\right)\right|+\left|G_j\left(\left\{v_j^x\right\}\right)\right|\cdot\left|G_j\left(\left\{v_j^y\right\}\right)\right|} \tag{1}$$

其中，$G_j\left(\left\{v_j^x\right\}\right)$ 表示属性值 v_j^x 对应的集合，$\left|G_j\left(\left\{v_j^x\right\}\right)\right|$ 表示属性值 v_j^x 出现的次数，因此 $1\le\left|G\left(\left\{v_j^x\right\}\right)\right|,\left|G\left(\left\{v_j^y\right\}\right)\right|\le m$，$2\le\left|G\left(\left\{v_j^x\right\}\right)\right|+\left|G\left(\left\{v_j^y\right\}\right)\right|\le m$，且 $\delta_j^{Ia}\in\left[1/3,m/\left(m+4\right)\right]$，$m$ 为信息表的规模。在表 1 中，$G_1\left(\left\{A_2\right\}\right)=\left\{u_2,u_3\right\}$，$\left|G_1\left(\left\{A_2\right\}\right)\right|=2$。$\delta_2^{Ia}\left(B_1,B_2\right)=\frac{2\times 2}{2+2+2\times 2}=0.5$。

Table 1. An example of information table
表 1. 信息表的例子

U \ A	a_1	a_2	a_3
u_1	A_1	B_1	C_1
u_2	A_2	B_1	C_1
u_3	A_2	B_2	C_2
u_4	A_3	B_3	C_2
u_5	A_4	B_3	C_3
u_6	A_4	B_2	C_3

2) 属性间耦合相似度(**IeASV**)

属性间耦合是指属性之间的相互影响。属性 a_j 的属性值 v_j^x (x 表示此属性值在 u_x 行)和 v_j^y 的相似度，是由 v_j^x 和 v_j^y 受其他属性 $a_k\left(k\neq j\right)$ 的影响来计算求得。

$$\delta_{j|k}^{I}\left(v_j^x,v_j^y,V_k\right)=\sum_{v_k\in\bigcap}\min\left\{P_{k|j}\left(\left\{v_k\right\}\middle|v_j^x\right),P_{k|j}\left(\left\{v_k\right\}\middle|v_j^y\right)\right\}$$

其中，$V_k\left(k\neq j\right)$ 是属性 a_k 的值，$P_{k|j}\left(\left\{v_k\right\}\middle|v_j^x\right)=\dfrac{\left|G_k\left(\left\{v_k\right\}\right)\cap G_j\left(\left\{v_j^x\right\}\right)\right|}{\left|G_j\left(\left\{v_j^x\right\}\right)\right|}$，而 $\left|G_j\left(\left\{v_j^x\right\}\right)\right|$ 表示属性值 v_j^x 出现的次数。$v_k\in\bigcap$ 表示 $v_k\in\varphi_{j\to k}\left(v_j^x\right)\bigcap\varphi_{j\to k}\left(v_j^y\right)$，而 $\varphi_{j\to k}\left(v_j^x\right)$ 表示与属性值 v_j^x 出现在同行的属性 a_k 的值的集合。

例如，计算 a_2 的两个属性值 B_1,B_2 受属性 a_1 影响下的相似度 $\delta_{2|1}^{I}\left(B_1,B_2,\left\{A_i\right\}_{i=1}^{4}\right)=\min\left\{P_{1|2}\left(\left\{A_2\right\}\middle|B_1\right),P_{1|2}\left(\left\{A_2\right\}\middle|B_2\right)\right\}=\min\left\{\dfrac{1}{2},\dfrac{1}{2}\right\}=0.5$。其中，$\varphi_{2\to1}\left(B_1\right)\bigcap\varphi_{2\to1}\left(B_2\right)=\left\{A_2\right\}$，$P_{1|2}\left(\left\{A_2\right\}\middle|B_1\right)=\dfrac{\left|G_1\left(\left\{A_2\right\}\right)\cap G_2\left(\left\{B_1\right\}\right)\right|}{\left|G_2\left(\left\{B_1\right\}\right)\right|}=\dfrac{\left|\left\{u_2\right\}\right|}{\left|\left\{u_1,u_2\right\}\right|}=\dfrac{1}{2}$。

属性 a_j 的属性值 v_j^x 和 v_j^y 的属性间耦合相似度(**IeASV**)被描述为：

$$\delta_j^{Ie}\left(v_j^x,v_j^y,\left\{V_k\right\}_{k\neq j}\right)=\sum_{k=1,k\neq j}^{n}\alpha_k\delta_{j|k}^{I}\left(v_j^x,v_j^y,V_k\right)\tag{2}$$

其中，$V_k\left(k\neq j\right)$ 是一组属性 a_k 的属性值，$k\in\left[1,n\right]$，n 为属性个数。α_k 是属性 α_k 的权重系数，$\sum_{k=1,k\neq j}^{n}\alpha_k=1,\alpha_k\in\left[0,1\right]$，并且 $\delta_j^{Ie}\in\left[0,1\right]$。例如在表 1 中，当 α_1和α_3 等于 0.5 时，$\delta_2^{Ie}\left(B_1,B_2,\left\{V_1,V_3\right\}\right)=0.5*\delta_{2|1}^{I}\left(B_1,B_2,\left\{A_i\right\}_{i=1}^{4}\right)+0.5*\delta_{2|3}^{I}\left(B_1,B_2,\left\{C_i\right\}_{i=1}^{3}\right)=0.25$。

综上，IaASV 强调的是属性内的属性值出现的频率，而 IeASV 强调的是属性间的相互影响。属性值耦合相似度(CASV)则通过这两种度量结合得到。

3) 属性值耦合相似度(**CASV**)

属性值 v_j^x 和 v_j^y 的耦合相似度(**CASV**)定义如下：

$$\delta_j^{A}\left(v_j^x,v_j^y,\left\{V_k\right\}_{k=1}^{n}\right)=\delta_j^{Ia}\left(v_j^x,v_j^y\right)\cdot\delta_j^{Ie}\left(v_j^x,v_j^y,\left\{V_k\right\}_{k\neq j}\right)\tag{3}$$

其中 δ_j^{Ia} 和 δ_j^{Ie} 分别是 IaASV 和 IeASV。在表 1 中，属性值 B_1 和 B_2 的 CASV 是

$$\delta_2^{A}\left(B_1,B_2,\left\{V_1,V_2,V_3\right\}\right)=\delta_2^{Ia}\left(B_1,B_2\right)*\delta_2^{Ie}\left(B_1,B_2,\left\{V_1,V_3\right\}\right)=0.5*0.25=0.125。$$

4) 对象耦合相似度(**CASO**)

对象耦合相似度(**CASO**)可以由属性值耦合相似度(**CASV**)累计得到。对象u_x和对象u_y的耦合相似度(**CASO**)被定义为：

$$\mathrm{CASO}\left(u_x,u_y\right)=\sum_{j=1}^{n}\delta_j^A\left(v_j^x,v_j^y,\{V_k\}_{k=1}^{n}\right) \tag{4}$$

其中δ_j^A是指 CASV，v_j^x,v_j^y分别指对象u_x,u_y在属性a_j上的值，并且$1\le x,y\le m,1\le j\le n$。

在表 1 中，对象u_2和u_3的耦合相似度$\mathrm{CASO}\left(u_2,u_3\right)=\sum_{j=1}^{3}\delta_j^A\left(v_j^2,v_j^3,\{V_k\}_{k=1}^{3}\right)=0.5+0.125+0.125=0.75$。

2.2.2. 变色龙(Chameleon)聚类算法

变色龙聚类算法是利用动态模型的层次聚类算法，可以发现同构、自然的簇。它采用动态建模的方法来确定一对簇之间的相似度，不依赖于一个静态的、用户提供的模型，能自动地适应被合并簇的内部特征。所以，当簇具有不同的形状、大小和密度时，变色龙算法也能够有效地聚类。

变色龙算法可分为构造 *k*-最近邻图，用最大生成树划分该 *k*-最邻近图和合并子簇形成最终簇三个阶段。在第三阶段，通过子簇C_i和C_j之间的相对互连性$RI\{C_i, C_j\}$和相对近似度$RC\{C_i, C_j\}$来决定两个子簇之间的相似度，然后选择使得相似度函数取值最大的两个子簇合并。

3. 基于耦合关系的医生用药异常分析

本文提出的框架(如图 1 所示)由三个阶段组成：定性分析，把现有的医院的真实处方转换成合适的表示形式，处方按相似团体进行划分，为下阶段定量分析做准备。定量分析，利用耦合关系度量处方间的相似性，用变色龙算法对处方进行聚类。异常检测，利用自定义的聚类指标检测医生用药异常。

3.1. 定性分析

为方便说明，先给出一些相关概念。

定义 1 (专向科室)：是指专门诊治某一类疾病或某一身体部位病症的科室。例如，眼科、皮肤科。

定义 2 (综合科室)：除了专向科室的科室均视为综合科室，它不具体诊治某一类疾病或某一身体部位病症。例如，简易门诊、急诊。

定义 3 (相似团体)：综合科室中相同疾病为一个相似团体，是跨门诊的；而专向科室中每一个科室中相同疾病为一个相似团体。

在相似团体内的处方中进行异常检测，才有意义，还可以减少计算时间，减少结果偏差。

3.2. 定量分析

定量分析由相似度计算和聚类两部分组成。相似度计算引入了耦合。考虑药物和药物之间的耦合关系，更真实地反应药物间的相互影响。如有些药品之间是有相互辅助作用的，它们往往同时出现。

3.2.1. 相似度计算

本文以处方为数据对象，药物为属性。为方便计算相似度，采用 1 或 0 表示此处方是否使用该药物。数据组织如图 2 所示，其中第一列代表处方编号。

利用信息表中属性及对象的耦合相似度度量方法(2.2.1节中已介绍)对处方数据进行相似度度量，其步骤如下：

1) 使用公式(1)计算属性值v_j^x和v_j^y的内耦合相似度(IaASV)；

2) 使用公式(2)计算属性值v_j^x和v_j^y的间耦合相似度(IeASV)；

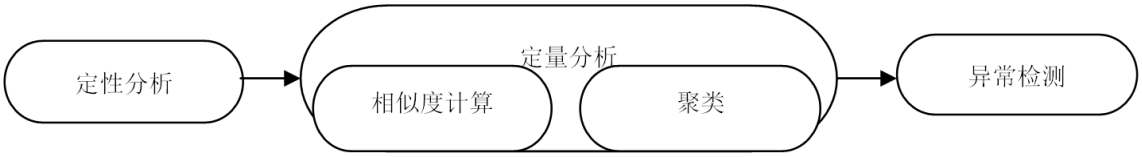

Figure 1. The framework for analyzing abnormal use of drugs
图 1. 医生用药异常分析框架

	1氯化钠注	1氨溴索注	1利巴韦林	1维生素C注	1%葡萄糖注	1小儿金翅	1氨溴特罗	1头孢硫脒	1青霉素钠	1炎琥宁注	1布洛芬缓
150	1	1	1	1	1	0	0	0	0	0	0
175	0	0	0	0	0	1	1	0	0	0	0
179	0	0	0	0	1	0	0	1	1	1	1
251	0	0	0	0	1	0	0	1	0	1	0
260	0	0	0	0	0	0	0	0	0	0	0
338	1	0	0	0	0	0	0	1	1	0	0
518	0	0	0	0	1	0	0	0	1	0	0
560	0	0	0	0	0	0	0	0	0	0	1
616	0	0	0	0	0	0	0	0	0	0	0
746	0	1	0	0	0	0	0	0	0	0	0
763	0	0	0	0	0	0	0	0	0	0	0
832	0	0	0	0	0	0	0	0	0	0	0
833	0	0	0	0	0	0	0	0	0	0	1
947	0	0	0	0	1	1	0	0	0	1	0
1022	0	0	0	0	1	0	0	1	0	1	0
1348	0	0	0	0	1	0	0	1	1	1	0
1378	1	1	0	0	1	0	0	0	0	1	0
1447	0	0	0	0	0	0	0	0	0	0	0
1452	0	0	0	0	1	0	0	0	0	0	0
1511	0	0	0	0	0	0	0	0	0	0	0
1576	0	0	0	0	0	0	0	1	0	1	0

Figure 2. Data organization
图 2. 数据组织

3) 使用公式(3)计算属性值 v_j^x 和 v_j^y 的耦合相似度(CASV)；

4) 使用公式(4)计算对象 u_x 和对象 u_y 的耦合相似度(CASO)。

计算 CASO 的时间复杂度为 $O\left(n^2R^3\right)$，n 为属性个数，R 为各属性中不同属性值个数的最大值，即：$R=\max_{j=1}^{n}\left\{V_j\right\}$。

下面举例说明 CASO 的计算过程，表 2 是由三个处方构成的实例，利用 CASO 计算得到各处方之间的相似度，如表 3 所示，表 2 中的处方之间越相似，表 3 中的相似度值越大。

3.2.2. 聚类

本论文采用变色龙算法进行聚类，如算法 1 所示，变色龙有两个关键步骤：图划分和层次聚类。

本文中一个处方表示一个数据点。若点 u_x 是点 u_y 的 k 最邻近(或点 u_y 是点 u_x 的 k 最邻近)，则两点之间存在一条边。变色龙算法第一阶段(第 1，2 步)是将处方聚类成若干子簇，每个子簇含有足够多的处方以满足动态建模生成子簇。将 k-最近邻图划分成几部分，极大减少割边。由于 k-最邻近图中的每条边表示处方之间的相似度，最小割边的划分可以有效减少不同子簇的处方间的关系。第二阶段(第 3~5 步)，使用动态建模架构层次聚类数据项，合并成簇。找到子簇之后，变色龙算法使用相对接近度和相对互连度框架对子簇进行合并。

假定 m 是数据对象(处方)的个数，p 是子簇的个数。划分图需要的时间复杂度[9]是 $O\left(mp+m\log_2 m\right)$ (第 2 步)，在图划分得到的 p 个子簇上进行凝聚层次聚类需要 $O\left(p^2\log_2 p\right)$ 时间(第 3~5 步)。

Table 2. Examples of description
表 2. 处方实例

处方编号	诊室	病患	医生	处方
1	儿科门诊	马**	刘**	1、氯化钠注射剂；2、氨溴索注射液；3、5%葡萄糖注射剂；4、炎琥宁注射剂；5、头孢硫脒注射剂(限儿童)
2	儿科门诊	党**	王*	1、布洛芬缓释混悬液；2、四季抗病毒合剂；3、肺力咳合剂(乙) (限儿童)；4、阿莫西林克拉维酸钾分散片
3	儿科门诊	叶**	邹*	1、布洛芬缓释混悬液；2、四季抗病毒合剂；3、肺力咳合剂(乙) (限儿童)；4、阿莫西林干混悬剂

Table 3. The CASO similarity between descriptions
表 3. 处方间 CASO 相似度

	1	2	3
1	0	0.218	0.119
2	0.218	0	0.873
3	0.119	0.873	0

Algorithm 1. Chameleon
算法 1. 变色龙

1：构造 *k*-最近邻图

2：使用多层图划分算法划分图

3：repeat

4：合并关于相对互连度和相对接近度而言，最好地保持簇的自相似性的簇

5：until 不再有可以合并的簇。

3.3. 异常检测

利用变色龙算法对处方聚类后，再对结果进行统计，将处方的聚类结果映射到医生。通过处方异常点率 *PO* 和簇内处方比率 *CP* 对医生用药进行异常检测，*PO* 和 *CP* 定义为：

$$PO = \frac{O_{t_i}}{C_{t_i}} \tag{5}$$

$$CP = \frac{N_{t_{ij}}}{N_{t_j}} \tag{6}$$

其中 O_{t_i} 为聚类后相似团体 t 中第 i 个医生的异常点处方数，C_{t_i} 为相似团体 t 中第 i 个医生的处方总量，$N_{t_{ij}}$ 为在相似团体 t 中，第 i 个医生在 j 簇中的处方数，N_{t_j} 为相似团体 t 中 j 簇包含的处方总数。

公式(6)主要检测：某医生诊治某相似团体次数较多，且使用的处方大致相似(医生一般都有自己的用药习惯)，会导致该医生的处方可能自成一簇。将自成簇的医生按异常报告出来，由领域专家鉴别。

在异常检测中，*PO* 和 *CP* 的大小分别与异常点率阈值 C_1 和簇内处方比率阈值 C_2 进行比较，如果 *PO* 或 *CP* 大于相应阈值，即可认定该医生在此相似团体中用药异常。其中，C_1 应大于该相似团体所有异常点数量与该相似团体处方总量的比值，C_2 的值需要考虑到该医生在此相似团体中处方比重较大的情况，因此 C_2 应大于该医生在此相似团体中的处方数量与该相似团体中总的处方数量的比值。

例如在数据规模为 500 的情况下，A 医生在综合诊室的上呼吸道感染这个相似团体中所开设处方 104 个，经聚类后该相似团体共有 62 个异常点，其中 A 医生共有 5 个异常点，将处方的结果映射到医生后，对 A 医生的聚类统计结果如表 4 所示。

按照上述阈值 C_1 和 C_2 的取值原则，C_1 大于 62/500 = 0.124，取 0.3，C_2 大于 104/500 = 0.208，取 0.4。A 医生的处方群点率为 5/104 = 0.048 小于 C_1。A 医生在簇 2 中处方比率为 60/113 = 0.53 大于 0. 4 ，因此可以把 A 医生在该相似团体的簇 2 中的用药情况报告出来，待专家鉴别。

4. 实验结果与分析

本论文中对象间耦合相似度(CASO)的计算由 MATLAB 2012b 完成，变色龙算法由 JAVA 语言实现。所有实验均在主频 2.4 GHz 的因特尔酷睿 2 处理器，1 GB 内存，windows7 操作系统的 PC 机上运行。

4.1. 数据集

本实验使用的是某医院的真实处方数据。本着实验数据全面、准确的原则，采用的是有多名医生诊治且每名医生都诊治较多次的相似团体——综合诊室的上呼吸道感染和消化内科的慢性胃炎两个相似团体，其中综合诊室的上呼吸道感染共涉及 93 个药品，经预处理后有 50 个药品，消化内科的慢性胃炎共涉及 49 个药品，即两类处方数据分别有 50 和 49 个属性，处方规模均为 1000。

4.2. Jaccard 和 CASO 对比分析

实验利用离散属性相似度计算中应用比较广泛的 Jaccard 相似度与对象耦合相似度 CASO 进行相似度度量，对比使用两者进行变色龙聚类得到的异常处方数和子簇个数，结果如图 3，图 4 所示。

从图 3 中可以看出，使用 Jaccard 的异常点明显多于使用 CASO 的异常点。且随着数据集规模的增加，使用 Jaccard 的异常点数量增长速度明显大于使用 CASO 的增长速度。但不论使用哪种方法，上呼吸道感染的异常点数量均大于慢性胃炎，这是因为慢性胃炎的治疗药物主要集中在少数几个药品中，且处方的重复率较高。从图 4 可以看出，使用 Jaccard 的簇数量多于使用 CASO 的簇数量。经领域专家对聚类结果进行鉴定，由 CASO 得到的簇和异常点更合理和准确。这是由于 Jaccard 仅仅考虑两个对象间同时出现的属性个数是完全不够的，还需要考虑属性间耦合作用和属性内耦合作用，比如氨溴索注射液和氨溴索注射剂是不可能在同一处方中出现的，CASO 就会考虑到这种情况。

从上述实验结果可以看出，CASO 计算处方间相似度时充分考虑了属性以及对象的耦合作用，在聚类时有较好的效果。但是，正因为 CASO 充分考虑了耦合作用，其计算时间复杂度较高，并且其时间复杂度受属性个数影响很大，3.1 节已阐述。

4.3. CASO 的计算时间随着属性数量增加的变化

本节对 CASO 的计算时间随着属性数增加的变化情况进行了实验。从图 5 可以看出随着属性数量的

Table 4. The clustering results of doctor A
表 4. A 医生聚类结果

	簇中的处方数量	该簇中 A 医生的处方数量
簇 1	54	9
簇 2	113	60
簇 3	87	8
簇 4	184	22

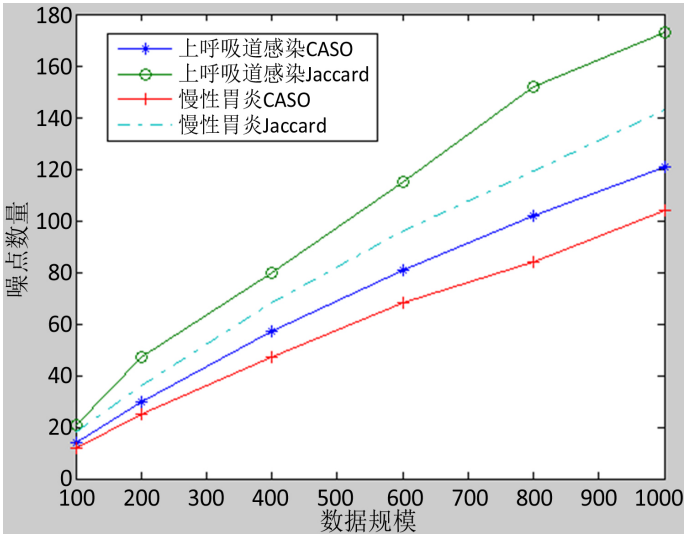

Figure 3. Influence of data size on the number of outliers
图 3. 数据规模增加对异常点数量的影响

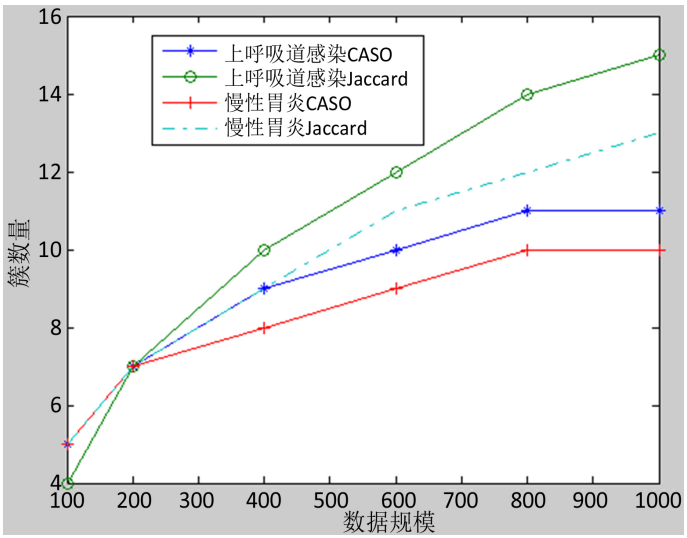

Figure 4. Influence of data size on the number of clusters
图 4. 数据规模增加对簇数的影响

增加，CASO 的计算时间增长较快。改善 CASO 计算复杂度也是未来工作中的主要任务之一。

4.4. 异常结果分析

将处方的聚类结果映射到医生，检验医生在该相似团体中用药是否异常。本文在消化科室的慢性胃炎的 1000 个处方中进行检验。表 5 为将处方聚类后的簇映射到医生时的聚类情况。而表 6 为每个医生的异常点数及其在该相似团体内的处方总数。

表 7 为异常检测结果，其中的异常情况列若为“异常点”，表示此医生所开的处方异常点比率(公式(5)所示)大于阈值 C_1。异常情况列若为“簇*、簇*”，表示此医生在这几个簇中的处方比重很大，被怀疑是自成簇。从表中可以看出医生 B、C、D、E、F、G 均有所异常，对簇中包含的主要药品进行分析可以发现，B 医生偏于使用的药品有莫沙必利分散片、瑞巴派特片、泮托拉唑钠肠溶胶囊；C 医生偏于使

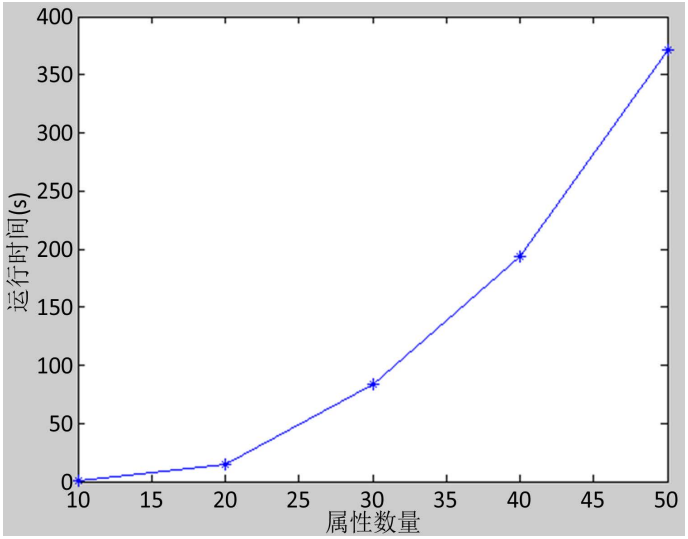

Figure 5. Influence of data size on the number of clusters
图 5. 数据规模增加对簇数的影响

Table 5. The clustering results after matching to doctors
表 5. 映射为医生后的聚类结果

	簇 1	簇 2	簇 3	簇 4	簇 5	簇 6	簇 7	簇 8	簇 9	簇 10
A 医生	48	0	14	0	0	13	2	2	0	3
B 医生	53	2	16	30	24	18	10	16	1	0
C 医生	52	0	4	11	4	3	1	33	2	3
D 医生	152	110	1	85	1	3	0	3	42	0
E 医生	11	0	1	0	4	4	1	5	0	1
F 医生	32	0	17	0	17	2	24	2	0	0
G 医生	2	0	0	0	0	0	0	0	0	11
总计	350	112	53	126	50	43	38	61	45	18

Table 6. The distribution of outliers after matching to doctors
表 6. 异常点映射为医生后情况表

	异常点数	处方总数
A 医生	4	86
B 医生	19	189
C 医生	10	123
D 医生	37	434
E 医生	14	41
F 医生	18	112
G 医生	2	15
慢性胃炎相似团体	104	1000

Table 7. The results of abnormal detection
表 7. 异常检测结果

	阈值 C_1	阈值 C_2	异常情况
A 医生	0.2	0.3	无
B 医生	0.2	0.3	簇 3、簇 5、簇 6
C 医生	0.2	0.3	簇 8
D 医生	0.2	0.6	簇 2、簇 4、簇 9
E 医生	0.2	0.2	异常点
F 医生	0.2	0.35	簇 7
G 医生	0.2	0.1	簇 10

用养蔚灵颗粒；D 医生则更加偏于使用胶体果胶铋干混悬剂、曲美布汀胶囊、猴头菌颗粒、复方尿囊素片、二氯乙酸二异丙胺片、伊托必利片；G 医生虽然关于慢性胃炎的处方较少，但大部分都集中在复方聚乙二醇电解质散；而 E 医生虽然因为异常点认定为在诊治慢性胃炎时用药异常，但并没有为病人使用固定药物。

5. 结束语

本文针对医患矛盾中的医生用药异常问题，提出一种医生用药异常检测的框架。该框架包括定性分析，定量分析和异常检测三部分。其中定量分析中集成了处方数据的耦合相似度度量方法和变色龙聚类算法。实验验证了所提出的框架能够较好的发现医生用药异常的情况。未来的工作是改进利用 CASO 计算处方耦合相似度的效率，以及在此框架下提出更高效的算法进行医生用药异常检测。

基金项目

国家自然科学基金(61472346, 61662086)；云南省自然科学基金(2016FA026, 2015FB149)。

参考文献 (References)

[1] 李璐瑒. 过度医疗——医生道德沦丧是表象医疗体制偏差是根本[J]. 首都医药, 2010(1): 20-21.

[2] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 著. 数据挖掘导论[M]. 第二版. 范明, 范宏建, 等, 译. 北京: 人民邮电出版社, 2011: 43-48.

[3] Lin, D. (1998) An Information-Theoretic Definition of Similarity. *ICML*, **1**, 296-304.

[4] Cai, D., He, X. and Han, J. (2005) Document Clustering Using Locality Preserving Indexing. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 1624-1637. https://doi.org/10.1109/TKDE.2005.198

[5] Figueiredo, F., Rocha, L., Couto, T., *et al*. (2011) Word Co-Occurrence Features for Text Classification. *Information Systems*, **36**, 843-858. https://doi.org/10.1016/j.is.2011.02.002

[6] Wang, C., Dong, X.J., Zhou, F. and Cao, L.B. (2015) Coupled Attribute Similarity Learning on Categorical Data. *IEEE Transactions on Neural Networks and Learning Systems*, **26**, 781-797. https://doi.org/10.1109/TNNLS.2014.2325872

[7] Wang, C., She, Z. and Cao, L.B. (2013) Coupled Attribute Analysis on Numerical Data. *Proceedings of the* 23*rd International Joint Conference on Artificial Intelligence* (*IJCAI* 2013), Beijing, 3-9 August 2013, 1736-1742.

[8] Dunn, J.C. (1974) Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, **4**, 95-104. https://doi.org/10.1080/01969727408546059

[9] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 著. 数据挖掘导论[M]. 第二版. 范明, 范宏建, 等, 译. 北京: 人民邮电出版社, 2011: 381-384.

[10] Marques, J.P., 著. 模式识别——原理方法及应用[M]. 第二版. 吴逸飞, 译. 北京: 清华大学出版社, 2002: 51-74.

[11] Chaturvedi, A.D., Green, P.E. and Carroll, J.D. (2001) K-Modes Clustering. *Journal of Classification*, **18**, 35-56. https://doi.org/10.1007/s00357-001-0004-3

[12] Huang, Z. (1998) Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, **2**, 283-304.

[13] Ding, C. and He, X. (2004) K-Nearest-Neighbor in Data Clustering: Incorporating Local Information into Global Optimization. *Proceedings of the ACM Symposium on Applied Computing*, Nicosia, 14-17 March 2004, 584-589.

[14] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of Second International Conference on Knowledge Discovery and Data Mining*, Portland, 2-4 August 1996, 226-231.

[15] Birant, D. and Kut, A. (2007) ST-DBSCAN: An Algorithm for Clustering Spatial-Temporal Data. *Data & Knowledge Engineering*, **60**, 208-221. https://doi.org/10.1016/j.datak.2006.01.013

[16] 张欣. 基于数据挖掘的纳税数据异常检测研究与应用[D]: [硕士学位论文]. 西安: 西安石油大学, 2009.

[17] Song, Y., Cao, L., Wu, X., *et al*. (2012) Coupled Behavior Analysis for Capturing Coupling Relationships in Group-Based Market Manipulations. *Proceedings of the* 18*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, 12-16 August 2012, 976-984. https://doi.org/10.1145/2339530.2339683

[18] 潘芳. 基于贝叶斯的防病患欺诈模型研究[J]. 现代商贸工业, 2014(10): 80-82.

[19] 王凯. 社保医疗消费中的异常信息检测研究[D]: [硕士学位论文]. 湖南: 中南林业科技大学, 2012.

[20] 梁俊, 孙听雪. 基于数据挖掘的标准化医疗保险监控模型构建[J]. 医学信息学杂志, 2015, 36(3): 42-46.