计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400062

面向病灶与其表征关联提取的核医学诊断文本挖掘

韩成成[1,2]　林　强[1,2]　满正行[1,2]　曹永春[1,2]　王海军[3]　王维兰[4]
1 西北民族大学数学与计算机科学学院　兰州 730030
2 西北民族大学动态流数据计算与应用实验室　兰州 730012
3 甘肃省人民医院核医学科　兰州 730020
4 西北民族大学国家教育部民族语言和信息技术重点实验室　兰州 730030
(2307115582@qq.com)

摘　要　医学影像是现代临床医学疾病诊治不可或缺的重要组成部分,SPECT是功能影像的主要成像技术,广泛应用于肿瘤骨转移等疾病的诊治。SPECT诊断报告文本包含患者个人信息、图像描述和建议性结果等几个方面的信息。为准确提取SPECT核医学骨显像诊断文本中疾病与其表征之间的关联关系,研究并提出基于数据挖掘的核医学文本关联规则挖掘方法。首先,针对核医学诊断文本可能包含的信息冗余、数据缺失及表述不一致等问题,提出SPECT核医学诊断文本的预处理及统一编码方法;然后,应用经典的关联规则挖掘算法Apriori,提出病灶与表征之间关联的挖掘算法;最后,使用一组源自三甲医院核医学科的真实SPECT核医学诊断文本数据,验证了所提出的方法。结果表明,提出的方法客观提取了疾病与其表征之间的关联,获得的客观性评价指标平均值不低于90%。

关键词:医学影像;SPECT核医学;诊断文本;文本挖掘;关系规则提取
中图法分类号　TP391

Mining Nuclear Medicine Diagnosis Text for Correlation Extraction Between Lesions and Their Representations

HAN Cheng-cheng[1,2],LIN Qiang[1,2],MAN Zheng-xing[1,2],CAO Yong-chun[1,2],WANG Hai-jun[3] and WANG Wei-lan[4]
1 School of Mathematics and Computer Science,Northwest Minzu University,Lanzhou 730030,China
2 Key Laboratory of Streaming Data Computing Technologies and Application,Northwest Minzu University,Lanzhou 730012,China
3 Department of Nuclear Medicine,Gansu Provincial Hospital,Lanzhou 730020,China
4 Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education,Northwest Minzu University,Lanzhou 730030,China

Abstract　Medical imaging is an indispensable part of the diagnosis and treatment of diseases in modern clinical medicine. SPECT is the main functional imaging technology and has been widely used in the diagnosis and treatment of diseases such as tumor bone metastasis. The SPECT diagnostic text contains several aspects of patients' personal information,image description,and suggested results. In order to accurately extract the association between disease and its representation in the diagnostic text of SPECT nuclear medicine bone imaging,a method of mining association rules of nuclear medicine text based on data mining is proposed. Firstly,a method of SPECT medical diagnostic text preprocessing and uniform coding is proposed to solve the problems of information redundancy,data loss and inconsistent expression. Secondly,the classical association rule mining algorithm Apriori is applied to propose the association mining algorithm between lesions and their representations. Finally,the proposed method is validated with a set of real-world SPECT nuclear medical diagnostic text data from the department of nuclear medicine in a 3a grade hospitals,and the results show that the proposed method is able to objectively extracted the association between the disease and its representation,and the average objectivity is more than 90%.

Keywords　Medical imaging,SPECT nuclear medicine,Diagnostic text,Text mining,Extraction of association rules

---

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:西北民族大学中央高校基本科研业务费专项资金资助研究生项目(Yxm2020101);国家自然科学基金项目(61562075);西北民族大学甘肃省一流学科引导专项资金(11080305);国家民委创新团队计划([2018]98)
This work was supported by the Northwest Minzu University for Central University Basic Scientific Research Operating Expenses Special Fund to Support the Graduate Program (Yxm2020101),National Natural Science Foundation of China (61562075),Gansu Provincial First-class Discipline Program of Northwest Minzu University (11080305) and Program for Innovative Research Team of SEAC ([2018] 98).
通信作者:林强(qiang.lin2010@hotmail.com)

1 引言

医学影像是现代临床医学疾病诊治不可或缺的重要组成部分，是以非侵入方式呈现机体内部结构与功能状态的重要技术手段。医学影像由医学成像系统（Medical imaging system）和医学图像处理（Medical image processing）两个相对独立的部分构成，其中医学图像处理以医学成像系统获取的图像为对象，研究图像的复原、增强、去噪、特征提取等预处理技术和图像模式分类、病灶分割、量化评价等图像理解技术。

自 1895 年英国物理学家伦琴发现 X 射线以来，医学成像从计算机断层扫描（Computerized Tomography，CT）、核磁共振（Magnetic Resonance Imaging，MRI）、超声波（Ultrasound）等传统结构成像模态发展到目前的单光子发射计算机断层成像术（Single-Photon Emission Computed Tomography，SPECT）和正电子发射断层成像术（Positron Emission Tomography，PET）等功能成像模型，进而催生了 SPECT/CT，SPECT/MRI，PET/CT 和 PET/MRI 混合成像模态。

不同于结构成像，功能成像需要预先在患者体内注入放射性药物（如 99Tmc-MIBI），经过一定时长的代谢后，用体外探头捕获体内特定部位或全身的药物残留分布，即放射量。大量临床验证表明，病变部位的放射量通常高于正常部位。这种依据放射量而非器官形态差异实现疾病检测的方法，有效弥补了传统结构成像的不足，因为有些疾病引发的器官或组织形态改变可能显著滞后于疾病可观察的症状。

恶性肿瘤骨转移是骨骼结构发病率最高的恶性肿瘤[1]，99mTc-亚甲基二磷酸盐（99mTc-MDP）全身平面骨显像已经广泛应用于骨转移疾病的诊断[2]。从疾病诊治的角度看，SPECT 核医学检查的输出结果是医疗工作人员手工生成诊断报告文本，包含患者个人信息、图像描述和建议性结果等几个方面的信息。

以诊断文本数据集为研究对象，探究患者与疾病之间、疾病与其表征之间的关联，是辅助医疗诊断研究的重要内容，在电子病例挖掘[3]、结构成像文本挖掘领域得到了初步探索。然而，目前尚未发现提取疾病与其表征关联关系的核医学诊断文本挖掘工作。

为准确提取 SPECT 核医学骨显像诊断文本中疾病与其表征之间的关联关系，本文研究并提出了基于数据挖掘的核医学文本关联规则挖掘方法。首先，针对核医学诊断文本可能包含的信息冗余、数据缺失及表述不一致等问题，提出 SPECT 核医学诊断文本的预处理及统一编码方法；然后，应用经典的关联规则挖掘算法 Apriori，提出病灶与表征之间关联的挖掘算法；最后，使用一组源自三甲医院核医学科的真实 SPECT 核医学诊断文本数据，验证了本文提出的方法。实验结果表明，本文提出的方法客观提取了疾病与其表征之间的关联，获得的客观性评价指标平均值不低于 90%。

2 相关工作

文本挖掘是数据挖掘的研究分支，主要包括数据准备、关系提取和数据挖掘 3 个步骤。其中，信息检索用于识别相关文本，信息抽取用于识别实体、关系等信息，数据挖掘则从结构化信息中识别出相互的关联[4-5]。医学文本挖掘已经成为学术界的研究热点，本研究就核医学挖掘工作，对研究现状分为两个方面进行介绍。

2.1 医学文本挖掘

以电子健康档案（Electronic Health Records，HERs）为研究对象，文献[3]研究了患者健康信息的提取并分析了疾病与健康信息之间的关系。因 EHRs 所包含的临床文本是半结构化的数据，增加了关系提取的难度，文献[6-7]综合应用文本分析、处理技术，将文本转换为结构化数值数据。Campbell 等[8]使用序列模式算法从大型电子健康记录数据集中挖掘儿童哮喘初始诊断的时间条件模式，用以揭示诊断之间未知的关联。针对年老患者也有类似的研究，McCoy 等[9]利用自然语言处理工具检查认知症状与偶发痴呆诊断之间的关系，有助于对痴呆风险进行分层。Yu 等[10]评估电子健康档案对注册老年护理家庭风险管理的贡献，其程度与老年护理认证所发挥的作用有关。越来越多的研究人员使用常规临床数据进行算法研究，Groenhof 等[11]评估了一个基于 EHRs 的数据挖掘算法的性能，并验证了算法的有效性。Rishi 等[12]使用自然语言理解技术诊断并试图提高急性心力衰竭患者的诊断效果。然而，EHRs 数据有着医学知识的抽象性，基于此，Liang 等[13]利用两种深度学习模型评估了计算机辅助医疗决策（Computer-Aided Medical Decision-Making，CAMDM）的有效性。Zhang 等统计了十多年的 EHRs 文献发布情况，从中可见我国电子健康领域已引起广泛关注，相关主题将成为研究者持续关注的领域[14]。

尽管同属文本挖掘的范畴，但本文研究疾病与其表征之间的关系，因而与上述工作的研究范畴有所不同。

2.2 传统结构影像

传统结构影像作为医学影像的鼻祖，有着广泛的适用性。CT 为检测新冠肺炎疾病的首选，Lei 等[15]分析了患者在新冠肺炎不同时期的临床表现。Liu[16]对比分析了 X 线和 CT 两种方法诊断下的患儿肺叶病变的主要特征，结果表明 CT 可降低漏诊率和误诊率。Weng 等[17]利用 CT 评估新型冠心病的诊断效果，结果表明该方法具有较高的准确度。Fei 等[18]研究了新生儿侵袭性真菌感染的临床特点与 MRI 影像学特征，实验结果表明临床与影像具有一定的联系。Zhu[19]探讨了对膝关节损伤进行 MRI 和 CT 检查的临床表现，通过对比分析发现 MRI 的诊断效果最佳。Zhao[20]观察了 CT 与 MRI 检查对小儿肝脏肿瘤的临床表现，对比发现 MRI 的诊断效果更好。Niu 等分析 MRI 联合超声在胎儿中枢神经系统畸形诊断中的临床价值，并为产前母儿保健检查内容选择提供参考[21]。

本文研究 SPECT 核医学文本中疾病与其表征之间的关系，聚焦疾病诊断知识库的构建，因而与现有研究的目标有所不同。

3 核医学诊断文本

3.1 核医学诊断文本数据

从疾病诊治的角度看，SPECT 核医学检查的输出结果是诊断报告文本，包含患者个人信息、图像描述和建议性结果等信息，表 1 给出了核医学诊断结果的示例。本文将患者的一次核医学检测结果称作一个病例（Medical case）。

表 1 SPECT 核医学文本样例

Table 1 Example of SPECT diagnosis text

SPECT 图像描述	诊断建议结果
静脉注射 99mTc-MDP 3 小时后行全身骨显像前位、后位：全身诸骨显影，双侧肘关节、腕关节、膝关节、踝关节及双手掌指关节见点片状放射性轻度浓聚，其余骨组织放射性呈左右对称分布未见明显异常浓聚灶及缺损区。双肾显影，形态未见异常。	双侧肘关节、腕关节、膝关节、踝关节及双手掌指关节炎。

病灶是疾病的具体表现，每个病灶都有其独特的表征。可以看出，表 1 左侧关于 SPECT 图像描述中的“双侧肘关节、腕关节、膝关节、踝关节及双手掌指关节”反映的是病灶所在的位置，“点片状”反映的是病灶的形状，“轻度”反映的是病灶的程度，而“浓聚”反映的是病灶的状态；右侧列诊断建议结果中“关节炎”反映的是诊断建议的疾病类别（简称类别）。

给定任意一个 SPECT 检查病例，总能找到位置、形状、程度、状态和类别信息（或者它们中的某几个），其中前 4 个是针对病灶的描述，最后一个是病灶归属疾病类别的表述。本文将描述病灶的这些信息称作病灶表征（Representation of Lesion），基于此，可将病灶形式化表示为如下五元组：

$$RL=(P,S,L,T,C) \tag{1}$$

其中，P 代表位置（Position）、S 代表形状（Shape）、L 代表程度（Level）、T 代表状态（sTate）、C 代表疾病类别（Class）。

3.2 诊断文本数据预处理

核医学诊断报告文本是医疗工作人员手工生成的自然语言文本，通过分析大量病例业已发现，信息遗漏、冗余、错误或表征信息缺失等现象在诊断报告中时有发生。例如，图 1 给出的 SPECT 图像描述中关于位置 P 的表述 L12，是一个明显的错误，因为人体脊柱中腰椎共有 5 块，标记为 L1－L5。

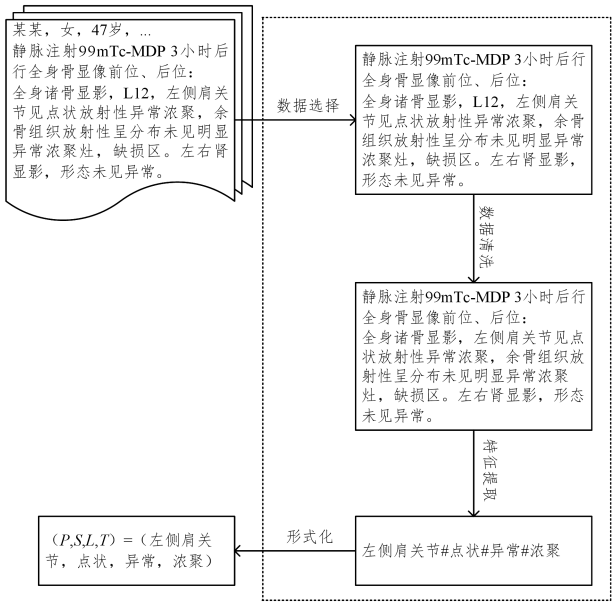


图 1 核医学诊断文本的预处理

Fig. 1 SPECT diagnosis text preprocessing

因此，首先需要对诊断报告文本进行预处理，以消除原始文本中包含的错误、冗余信息，同时补全遗漏的信息。图 1 给出了 SPECT 诊断报告文本的预处理过程。

首先，数据选择阶段隐去患者的姓名、性别、年龄等隐私信息，以尽可能做到隐私保护；其次，数据清洗阶段消除文本中的错误表述和冗余信息，同时结合上下文信息补全遗漏的信息；然后，表征提取阶段提取出文本中包含的位置、形状、状态和程度信息；最后，形式化阶段将提取出的位置、形状、状态和程度信息连同诊断结果的疾病类别信息，组成前述定义的病灶五元组。

需要注意的是，如何从 SPECT 诊断报告文本中高效准确地提取病灶的表征信息，即表征提取阶段的实现，是核医学文本挖掘的重要内容之一，归属自然语言理解的研究范畴。本文利用人工提取的病灶表征研究病灶各表征之间的关联。

4 病灶表征间关联的挖掘

从统计意义上看，病灶各表征信息之间必然存在某一种或几种关联，例如：有些疾病常发作于上肢，而另一些可能常发作于下肢；有些疾病的病灶总是呈现为条块状，而另一些却呈现为团状。为了探究特定疾病在病灶表征之间具有的特定关联，本文提出了基于数据挖掘的疾病表征关联规则提取方法，该方法由病灶表征的形式化编码和关联规则提取算法两部分构成。

4.1 病灶表征的形式编码

在病灶的五元组表示中，位置 P 实际上代表人体的骨骼。图 2 给出了人体骨骼系统的层次结构。可以看出，人体的骨骼具有明显的层次包含关系，因此，关联关系的挖掘处理要将这种关联关系纳入考虑，以从位置的不同层级上探究疾病的发作模式。

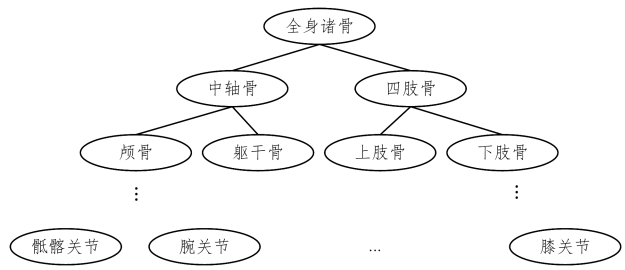
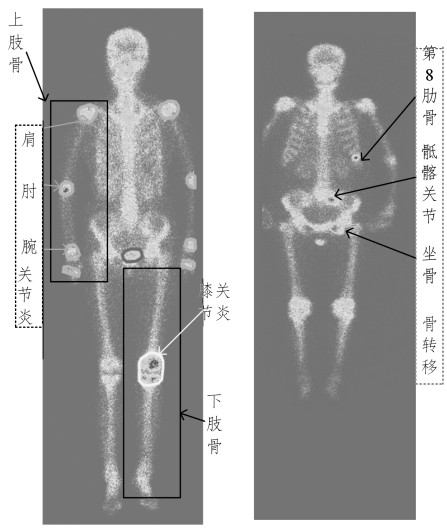


图 2 人体骨骼层次结构

Fig. 2 Hierarchical structure of humman skeleton

图 3 给出了两个以 RGB 格式呈现的 SPECT 核医学检查病例。



(a)关节炎病例 (b)骨转移病例

图 3 SPECT 图像

Fig. 3 SPECT images

可以看出，无论关节炎还是骨转移，病灶所在区域的像素值明显高于其他区域。这是因为，SPECT 核医学成像通过捕获注射药物在体内的残留实现疾病的检查，体内正常部位的药物会随着时间推移而逐渐代谢殆尽，而病变部分的药物残留能够被探头捕获。具体而言，关节炎病例中上右肢骨和下左肢骨均发生明显病变；骨转移病例中第 8 肋骨、骶髂关节以及坐骨均发生病变。

表 2 列出了 SPECT 骨显像诊断文本病灶的形状、状态、程度表征信息的全集，从中可以看出，每个表征的取值相互之间不存在包含关系。

表 2　病灶表征 P，S 和 L 及类别 C 的取值集合

Table 2　Values of P, S, L and C

表征	取值	数量
形状 S	块状、点状、片状、条片状、条块状、点块状、点条片状、点条状、片状	9
状态 T	浓聚、增强、聚集、减淡、稀疏、缺如、缺损、扩张、滞留、摄取、不规则、不均匀、清晰、形态失常、形态欠佳、边界欠规整、体积增大、体积缩小、肿胀、畸形	20
程度 L	略微、轻度、确定、明显、较强、高度、过度	7
类别 C	关节炎、骨转移、退行性改变、其他	4

结合图 2 和表 2，本文提出固定长度的病灶表征二进制度形式化编码如下：

位置 P：采用 21 位二进制编码，由于人体骨骼共分为 7 层（见图 2），所以获得图 4 所示的编码。

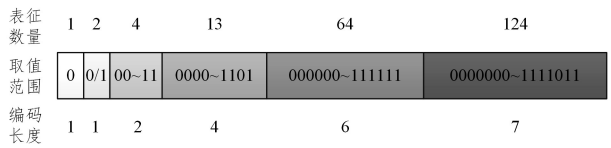


图 4　位置 P 的层次编码

Fig. 4　Hierarchical code for position P

形状 S：共 9 个，采用 4 位二进制编码。

状态 T：共 20 个，采用 5 位二进制编码。

程度 L：共 7 个，采用 3 位二进制编码。

类别 C：共 4 个，采用 2 位二进制编码。

上述编码规则将产生长度为 35 位的固定长度编码，如果诊断文本中的位置不出现在 P 的编码中，则用 * 填充，以确保编码长度固定。例如，参照图 3，按照位置 P 的二进制固定编码规则，肩关节对应的 21 位编码为：01100111100011 ******* 。表 3 列出了位置的编码示例。

应用上述形式化编码，挖掘算法可顺序扫描固定长度的编码，以方便抽取其中包括的不同表征值。此外，位置 P 的层级编码反映了位置的层级包含关系，有助于挖掘算法实现伸缩挖掘，即提取不同层级位置与疾病之间的关联。

表 3　位置编码示例

Table 3　Example of position code

位置	21 位固定长度编码
骶髂关节	000100110001101000001
腕关节	011010010110101000100
…	…
膝关节	011101011100111110111
肩关节	01100111100011 *******

4.2　基于 Apriori 的病灶-表征关联挖掘

Apriori 算法是经典的关联规则挖掘算法，有着结构简单及在小数据集上性能较好的优点。因此，本文提出了基于 Apriori 的病灶表征关联规则挖掘算法。

Apriori 算法包括频繁项集的产生和关联规则的产生两个主要过程，其中频繁项集的产生过程是查找那些满足支持度要求的项集，而关联规则是由满足最小置信度的频繁项集组成的集合。

本文病灶表征关联规则挖掘的目的在于探究特定疾病通常在什么位置发病、通常具有什么样的形状、呈现什么状态，以及表现出怎样的严重程度，进而形成特定的模式，以便构建基于知识的疾病诊断模型。

对于给定的由 n 个病灶五元组构成的集合 $RL=\{RL_1, RL_2, \cdots, RL_n\}$ 以及第 i 个病灶的表征 rl_1 和 rl_2（$rl_1, rl_2 \in RL_i$），定义支持度如下：

$$Supp(rl_1, rl_2)=P(rl_1 \cap rl_2)=\frac{1}{n-1}Count(RL_{1\leqslant j\neq i\leqslant n} \mid rl_1, rl_2 \in RL_j) \tag{2}$$

其中，函数 $Count(\cdot)$ 用于统计所有同时包含 rl_1 和 rl_2 的病灶的数量。

令 $Supp_{\min}$ 代表最小支持度阈值，若将表征 rl_1 和 rl_2 的同时发生视作事件，则称该事件是频繁的当且仅当 $Supp(rl_1, rl_2) \geqslant Supp_{\min}$。

本文的目的在于，就给定的支持度阈值，发现那些与疾病类别 C 同时发生的频繁事件。

不同于支持度，置信度是另一个关联规则的度量指标。类似地，可定义置信度如下：

$$Conf(rl_1, rl_2)=P(rl_1 \cap rl_2)/P(rl_1)=\frac{Count(RL_{1\leqslant j\neq i\leqslant n} \mid rl_1, rl_2 \in RL_j)}{Count(RL_{1\leqslant k\leqslant n} \mid rl_1 \in RL_k)} \tag{3}$$

同理可定义最小置信度阈值 $Conf_{\min}$。对于设定的最小支持度 $Supp_{\min}$ 和最小置信度 $Conf_{\min}$，本文提出的算法首先产生病灶表征的频繁项集，即统计每个表征在病例（表征）集中出现的次数，将那些满足最小支持度 $Supp_{min}$ 条件的病例组织起来形成 1 项集 S_1。然后，算法逐层递归扫描病例集，直至没有更大项集产生。

上述过程中最重要的步骤是最大项集 S_k 的产生，此过程需要通过将 S_{k-1} 与自身连接产生候选 S_k 集合。假定 s_1 和 s_2 是 S_{k-1} 的成员，用 $l_i[j]$ 表示 l_i 中的第 j 项。Apriori 算法按照患者的 ID 对数据进行排列，对于 $k-1$ 项集 s_i，排序得到 $s_i[1]<s_i[2]<\cdots<s_i[k-1]$。执行连接操作 S_{k-1} 与自身连接，如果 $(s_1[1]=s_2[1])$ && $(s_1[2]=s_2[2])$ && … && $(s_1[k-2]=s_2[k-2])$ && $(s_1[k-1]<s_2[k-1])$，则认为 s_1 和 s_2 是可连接的。连接 s_1 和 s_2 的结果是 $\{s_1[1], s_1[2], \cdots, s_1[k-1], s_2[k-1]\}$，在此过程中要确保不产生重复项。

例如，若项集 S_1 为{肩关节}，{膝关节}和{踝关节}，可生成 S_2 项集为{肩关节，膝关节}、{肩关节，踝关节}和{膝关节，踝关节}，按照此规则直到不能再找到频繁 k 项集为止。

根据以上过程得到的频繁项集，产生关联规则如下：

1）对于每个频繁项集 S，产生 S 的所有非空子集。

2）对于 S 的每个非空子集 U，若 $SupportCount(s_i \cap s_j)/SupportCount(U) \geqslant Conf_{\min}$，则输出规则 $S \Rightarrow (L-S)$。其中，$SupportCount$ 是项集的支持数量，$Conf_{\min}$ 是人为设定的最小置信度。具体如算法 1 所示。

算法 1 基于 Apriori 的病灶表征关联挖掘算法

Inputs:表征数据库 D;最小支持度 $Supp_{min}$;最小置信度 $Conf_{min}$

Outputs:频繁项集 L;关联规则

1. L_1=find_frequent_1_itemsets(D);
2. for
3. C_k=aproiri_gen(L_{k-1},$Supp_{min}$);
4. C_t=subset(C_k,t);
5. endfor
6. for
7. L_k=$\{c \in C_k | c.count \geqslant Supp_{min}\}$
8. end
9. return L
10. for
11. SubItems=GenSubItemSet(L_k);
12. AR_gen=AssociationRule(L_k,$Conf_{min}$);
13. end

算法 1 通过样本集合,生成含 k 个元素的候选项集(步骤 3),从而生成包含在事务 t 的候选项集 C_t,其中满足支持度阈值的条件时,生成病症及其表征的频繁项集 L_k(步骤 7)。频繁项集在候选项集中满足一定支持度阈值,从而生成关联规则。

5 实验与结果

应用一组在核医学临床检查中获取的 SPECT 诊断真实文本数据,本节验证本文提出的挖掘算法。

5.1 实验设计

本文实验数据来自三甲医院核医学科的 SPECT 检查报告文本,共包含 3830 个病例。因可能存在随访现象,所以实际包含的患者数少于病例数。

表 4 列出了实验数据涉及的患者年龄分布,可以看出,患者主要集中在 50～69 岁之间,而且以 60～69 岁之间居多。患者中女性占 53%,男性占 47%。

表 4 患者的年龄分布

Table 4 Age distribution of patients

年龄段/年	数量/人
10～19	44
20～29	103
30～39	153
40～49	645
50～59	991
60～69	1076
70～79	547
80～89	162
90～99	7

图 5 给出了本文实验数据涉及的疾病类型,对应于病灶五元组中的类别 C,其中关节炎 817 条,退行性改变 569 条,骨转移 644 条,其他疾病 1800 条。

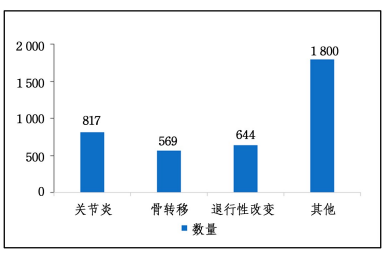


图 5 实验涉及的疾病类别

Fig. 5 Category of diseases involved in experiment

本文实验的评价指标是关联规则挖掘算法产生的结果的客观性,即算法挖掘结果是否与医生的专家知识一致。然而,客观性是一个非量化指标,不能很好地应用于实验结果的评价。对此,本文采取了如下半量化解决方案:

(1)量化评分。邀请 3 名核医学科领域专家,各自对照原始 SPECT 诊断病例对算法挖掘结果量化打分(1 表示完全一致;0.8 表示较一致;0.5 表示基本一致;0 表示不一致)。

(2)投票表决。针对 3 位领域专家的打分结果,采取少数服从多数和就低原则表决产生最终评分。

应用上述方案,每一条关联规则均会获得一个量化评价结果。

5.2 实验结果

为了考查不同支持度、置信度阈值对关联规则挖掘结果的影响,实验设置了 3 种不同的支持度、置信度取值方案,如表 5 所列。

表 5 支持度和置信度的不同取值

Table 5 Various values for *Supp* and *Conf*

方案	支持度 *Supp*	置信度 *Conf*
1	0.11	0.2
2	0.12	0.3
3	0.18	0.5

实验结果表明,当支持度设为 0.12,置信度设为 0.3 时,算法获得的关联规则具有较高的客观性。表 6 列出了位置 P 与疾病类别 C 之间的关联情况,其中,01 代表关节炎、10 代表退行性改变、11 代表骨转移、00000100000111001110 1 代表 L4、0000010000011100111 10 代表 L5、01111010110000 ******** 代表股骨,其余编码的含义如表 6 所列。

表 6 当 $Supp$=0.12,$Conf$=0.3 时位置 P 和类别 C 的关联规则

Table 6 Association rules between P and C when $Supp$=0.12 and $Conf$=0.3

关联规则	(支持度/%,置信度/%)
01110101110011111011→ 01	(21.14,83.80)
01 → 01110101110011111 0111	(21.14,77.66)
0000010000011100111 10 → 10	(18.44,75.51)
10 → 0000010000011100111 10	(18.44,52.56)
0000010000011100111 01 → 10	(13.96,76.08)
10 → 0000010000011100111 01	(13.96,39.77)
01111010110000 ******** → 11	(13.96,96.00)
11 → 01111010110000 ********	(13.96,37.23)
01100111100011 ******** → 01	(12.06,67.98)
01 → 01100111100011 ********	(12.06,44.32)

由表 6 列出的位置及其疾病的关联规则可以看出,不同的位置对应着不同的疾病,即不同的疾病发作的位置通常各不一样。

如前所述,位置按图 2 所示的层次方式编码,泛化位置(如膝关节)和具体位置(如左侧膝关节)分别与疾病的关联结果也有所不用。由表 7 可知,泛化位置与疾病的关联效果相对而言更好。其中,01110101110011111011 10 代表左侧膝关节、01110101110011111011 11 代表右侧膝关节、01110101100111111011 11 代表膝关节、01 代表关节炎。

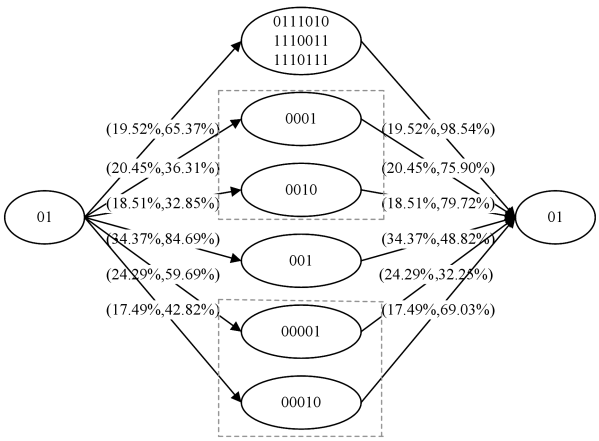
当支持度阈值设为 0.12、置信度阈值设为 0.3 时,图 6 给出了关节炎、骨转移、退行性改变与其表征之间的关联,其中,0001 代表点状,0010 代表片状,0100 代表点片状,00001 代表浓聚,00010 代表增强,001 代表轻度,010 代表异常,括

号内数值表示支持度和置信度。

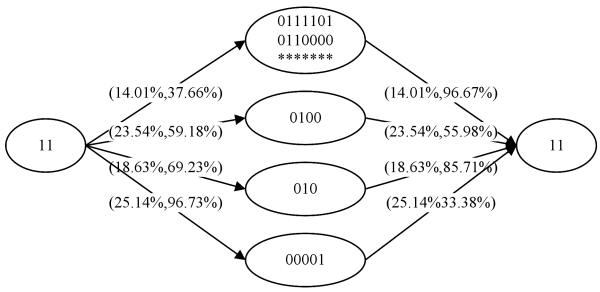
表 7 泛化位置及具体位置与疾病的关联情况

Table 7 Association rules between general and special positions and diseases

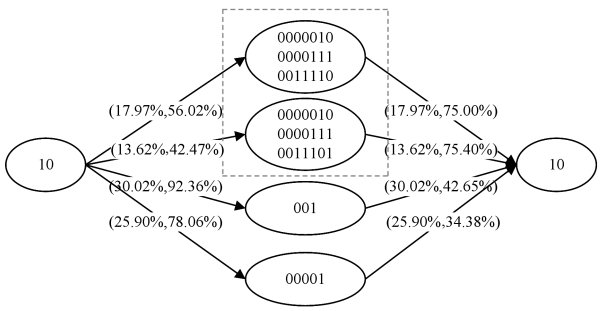
关联规则	(支持度/%，置信度/%)
01110101110011111101110 → 01	(18.52,90.26)
01 → 01110101110011111101110	(18.52,58.86)
01110101110011111101111 → 01	(18.11,89.11)
01 →01110101110011111101111	(18.11,57.53)
01110101110011111101111 → 01	(19.52,98.54)
01 → 01110101110011111101111	(19.52,65.37)



(a)关节炎



(b)骨转移



(c)退行性改变

图 6 3 类疾病与其表征之间的关联

Fig. 6 Association rules between three diseases and their presentations

由图 6 所示的 3 种疾病及其属性之间的关联可以看出，关节炎经常发生在膝关节，见点状/片状且轻度浓聚/增强。此外，由关节炎与其表征之间的支持度和置信度，可推断出当膝关节见点状/片状轻度浓聚/增强时，临床通常考虑此位置发生了关节炎病变。根据支持度与置信度的取值，也可进一步得知关节炎的病灶形状更可能是点状，病灶的状态更可能是浓聚。

类似地，骨转移经常发生在股骨，见点片状且异常浓聚，而当股骨见点片状异常浓聚时，更有可能发生骨转移。退行性改变经常发生在 L4(第 4 腰椎)和 L5(第 5 腰椎)，呈现轻度浓聚，对应可以理解为当 L4 和 L5 呈轻度浓聚时，发生退行性改变的可能性较大。

图 7 给出了 3728 条病例的客观性量化评价结果，从中可以看出，本文提出的挖掘算法的结果与临床诊断的结果保持高度一致，能够客观反映病灶各表征之间的关联，为基于知识的自动化诊断模型建构奠定了基础。本实验关节炎及其表征的关联结果具有实际参考情况，双腿支撑着人体骨骼，确保人的站立、行走等基本姿势，久而久之双腿势必会受到影响，评价结果比较高。而骨转移是其他癌症发生病变的结果，对于实际诊断文本来说，其中一部分病例是全身发生骨转移，因此评价结果相对来说要低一些。退行性改变严格来说并不属于疾病范畴，它只是随着年龄的增长，人体的各种器官的退化，对于实际诊断文本来说，多数人会在骨骼上发生退行性改变。

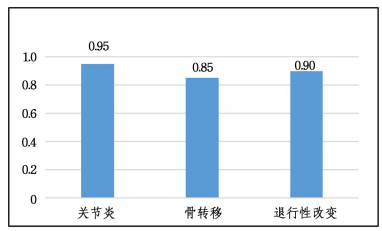


图 7 算法客观性量化结果

Fig. 7 Quantitative objectivity of proposed algorithm

尽管本文算法获得了较为客观的结果，但支持度和置信度阈值明显偏低。这是因为本文实验数据集中涉及的疾病类型多，而每一类包含的病例数明显偏少；另外，类不平衡问题明显存在。

结束语 以疾病的病灶与其表征之间的关联关系提取为目标，本文研究了基于数据挖掘的核医学文本关联规则方法，具体包括：1)提出了核医学文本的预处理方法，对核医学文本可能存在的信息遗漏、冗余、错误或表征信息缺失等进行了处理；2)提出了病灶表征的编码规则，将一条诊断文本编码为具有固定长度的二制度位串；3)提出了基于 Aprior 算法的病灶表征与所属疾病之间关联的挖掘算法，特别是算法考虑了位置的层次关系，因而确保算法的伸缩性；4)应用源自核医学科病例检测的真实数据，测试并验证了本文提出的方法。实验结果表明，该方法能够有效提取出病灶表征及其所属疾病之间的关联，获得的客观性评价指标的平均值不低于 90%。

基于本文工作，未来将从如下几个方法进一步开展研究：

1)拓展数据集，以研究更大样本集和更多疾病类别核医学文本的关联规则挖掘；

2)进一步融合领域知识，深入探究疾病表征的范围和量度；

3)研究并提出更加高效的关联规则挖掘算法，以实现在更大数据集上挖掘病灶表征与疾病类别间的关联；

4)应用本文研究成果，构建基于知识的核医学诊断规则和模型，以辅助核医学临床实践。

参 考 文 献

[1] VASSILIOU V，ANDREOPOULOS D，FRANGOS S，et al. Bone metastases：assessment of therapeutic response through radiological and nuclear medicine imaging modalities[J]. Clinical

Oncology (Royal College of Radiologists), 2011, 23(9): 632-645.
[2] ABIKHZER G, GOUREVICH K, KAGNA O, et al. Whole-body bone SPECT in breast cancer patients: the future bone scan protocol[J]. Nuclear Medicine Communications, 2016, 37(3): 247-253.
[3] REÁTEGUI R, RATTÉ S. Analysis of Medical Documents with Text Mining and Association Rule Mining[C]// International Conference on Information Technology and Systems. Springer, Cham, 2019, 1: 744-753.
[4] COHEN A M, HERSH W R. A survey of current work in biomedical text mining[J]. Briefings in Bioinformatics, 2005, 6(1): 57-71.
[5] WANG H C, ZHAO T J. Research and development of biomedical text mining technology[J]. Chinese Journal of Information, 2008, 22(3): 89-98.
[6] MINER G, ELDER IV J, FAST A, et al. Practical text mining and statistical analysis for non-structured text data applications [M]. Boston: Academic Press, 2012.
[7] WEISS S M, INDURKHYA N, ZHANG T, et al. Text mining: predictive methods for analyzing unstructured information[M]. Berlin: Springer Science & Business Media, 2010.
[8] CAMPBELL E A, BASS E J, MASINO A J. Temporal condition pattern mining in large, sparse electronic health record data: A case study in characterizing pediatric asthma[J]. Journal of the American Medical Informatics Association, 2020, 27(4): 558-566.
[9] MCCOY T H J, HAN L, PELLEGRINI A M, et al. Stratifying risk for dementia onset using large-scale electronic health record data: A retrospective cohort study[J]. Alzheimer's and Dementia: the Journal of the Alzheimer's Association, 2020, 16(3): 531-540.
[10] YU P, JIANG T, HAILEY D, et al. The contribution of electronic health records to risk management through accreditation of residential aged care homes in Australia[J]. BMC Medical Informatics and Decision Making, 2020, 20(1): 58.
[11] GROENHOF T K J, KOERS L R, BLASSE E, et al. Data mining information from electronic health records produced high yield and accuracy for current smoking status[J]. Journal of Clinical Epidemiology, 2020, 118: 100-106.
[12] RISHI V P, THIDA C T, SUE H S, et al. Can Natural Language Processing Improve the Accuracy of Identifying Acute Heart Failure in Electronic Health Records[J]. Circulation, 2018, 138(138): 16034.
[13] LIANG Z H, LIU J, OU A H, et al. Deep generative learning for automated EHR diagnosis of traditional Chinese medicine[J]. Computer Methods and Programs in Biomedicine, 2019, 174: 17-23.
[14] ZHANG K, WANG W T, XIE Y Q. Research progress of electronic health in China [J]. Library Forum, 2018, 38(8): 84-92.
[15] LEI Z Q, SHI H S, LIANG B, et al. Imaging detection and infection prevention and control of novel coronavirus (2019-nCoV) pneumonia [J]. Journal of Clinical Radiology, 2020, 39(1): 12-16.
[16] LIU J W. Application of computed tomography in diagnosis of mycoplasma pneumoniae pneumonia in children [J]. Chinese Convalesce Medicine, 2019, 28(3): 280-282.
[17] WENG T W, MAO D B, JIN J, et al. Computed tomography (ct) scan media stored blood flow to the score research progress [J]. Journal of Geriatric Medicine and Health Care, 2019, 5: 685-688.
[18] FEI Z H, PAN H P, LUO Z Q, et al. Clinical characteristics and magnetic resonance imaging of invasive fungal infection in neonates [J]. Journal of Chinese Hospital Infectious Diseases, 2019, 19: 161-165.
[19] ZHU H W. Effect observation of mri in diagnosis of knee joint injury [J]. Imaging Research and Medical Application, 2019, 3(15): 167-168.
[20] ZHAO N N. Comparison of ct and mri in pediatric liver tumors [J]. Tumor Foundation and Clinical, 2019, 6: 540-541.
[21] NIU Y Y, WANG Y M, CHEN N. Clinical application of nuclear magnetic resonance combined ultrasound in fetal central nervous system malformation [J]. Contemporary Medicine, 2020, 9(26): 90-92.

HAN Cheng-cheng, born in 1994, postgraduate, is a member of China Computer Federation. Her main research interests include data mining and intelligent information processing.

LIN Qiang, born in 1979, Ph.D, associate professor, master's supervisor, is a member of China Computer Federation. His main research interests include medical image computing, data stream mining, pervasive computing and intelligent information processing.