Journal of Biomedical Informatics 48 (2014) 73–83

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Inferring characteristic phenotypes via class association rule mining in the bone dysplasia domain

Razan Paul [a], Tudor Groza [a,*], Jane Hunter [a], Andreas Zankl [b,c]

[a] School of ITEE, The University of Queensland, Australia
[b] Bone Dysplasia Research Group, UQ Centre for Clinical Research (UQCCR), The University of Queensland, Australia
[c] Genetic Health Queensland, Royal Brisbane and Women's Hospital, Herston, Australia

ARTICLE INFO

*Article history:*
Received 6 August 2013
Accepted 1 December 2013
Available online 10 December 2013

*Keywords:*
Class association rule mining
Mining characteristic phenotypes
Bone dysplasias

ABSTRACT

Finding, capturing and describing characteristic features represents a key aspect in disorder definition, diagnosis and management. This process is particularly challenging in the case of rare disorders, due to the sparse nature of data and expertise. From a computational perspective, finding characteristic features is associated with some additional major challenges, such as formulating a computationally tractable definition, devising appropriate inference algorithms or defining sound validation mechanisms. In this paper we aim to deal with each of these problems in the context provided by the skeletal dysplasia domain. We propose a clear definition for characteristic phenotypes, we experiment with a novel, class association rule mining algorithm and we discuss our lessons learned from both an automatic and human-based validation of our approach.



## 1. Introduction

Finding, capturing and describing characteristic features (or symptoms) represents a key aspect in disorder definition, diagnosis and management. In general, such features are directly recognised by experts via repeated observations in patient cases. On the other hand, when the disorders are very similar, and share most of their phenome space, determining discriminative features is done in a pair-wise differential manner. This process is particularly important for rare disorders, as it may provide an initial screening and diagnosis direction, which could then prove to be vital. However, the sparse nature of the phenome space in rare disorders, and the limited number of experts makes this process very difficult. Hence, identifying characteristic features from existing patient cases in a (semi-) automatic manner would be highly beneficial for improving the understanding of and the shared agreement on the definition and characterisation of rare disorders.

From a computational perspective, these features raise two major challenges: (i) defining them in a computationally tractable way and (ii) devising appropriate algorithms to infer them, by exploiting their sparse nature. An additional, orthogonal, challenge is defining a sound validation mechanism that takes into account both the computational definition as well as the human expert opinion. In this paper, we describe our experiments and lessons learned from inferring characteristic features in the bone dysplasia domain.

Skeletal dysplasias [1] are a heterogeneous group of genetic disorders affecting skeletal development. Currently, there are over 450 recognised such disorders, structured in 40 groups. Patients with bone dysplasias have complex medical issues including skeletal deformations, impaired development and neurological complications. Since most skeletal dysplasias are very rare (<1:10,000 births), data on clinical presentation, natural history and best management practices is sparse. Another reason for data sparseness is clinical variability, i.e., the small number of clinical features typically exhibited by patients from the large range of possible phenotypic and radiographic characteristics usually associated with these disorders. Due to the rarity of these conditions and the lack of mature domain knowledge, correct diagnosis is often very difficult. In addition, only a few centres worldwide have expertise in the diagnosis and management of these disorders.

Different research groups around the world have, over time, built small patient registries that are neither open nor interoperable. In 2002, the European Skeletal Dysplasia Network (ESDN, http://www.esdn.org/) was created to alleviate, at least partly, the data sparseness issue, and at the same time to provide a collaborative environment to help with the diagnosis of skeletal dysplasias and to improve the information exchange between researchers. To date, ESDN has gathered over 1200 patient cases, which have been discussed by its panel of experts.

We have used the data acquired by ESDN to study a set of bone dysplasias with the above-mentioned goal of designing an

* Corresponding author.
*E-mail addresses:* razan.paul@uq.edu.au (R. Paul), tudor.groza@uq.edu.au (T. Groza), jane@itee.uq.edu.au (J. Hunter), a.zankl@uq.edu.au (A. Zankl).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.12.001

approach to automatically infer characteristic phenotypes. The high degree of subjectivity makes the understanding and capturing of the attributes that define such phenotypes problematic even for human experts. Hence, in order to provide a computationally tractable definition for them, we have considered a characteristic feature to be one that is (i) frequent for the disorder under scrutiny, i.e., its absence would rule out the current disorder and (ii) rare in other closely-related disorders, i.e., specific or discriminative for the current disorder. As a side remark, a feature is called pathognomonic for a disease if it identifies that disease beyond any doubt. Our ultimate aim is to find the set of features that come as close as possible to being pathognomonic. Another way of looking at characteristic features is by providing them a probabilistic interpretation of the form: the presence of feature F increases the probability of disorder D, or if F then D is more likely. Taking this probabilistic interpretation a step further allows us to map the process of inferring characteristic features to the problem of discovering class associations in the data mining field [2–4].

Association rules [5] provide knowledge in the form of probabilistic "if-then" statements. The head of the association rule (i.e., the if part) is called antecedent, while the body (i.e., the then" part) is called consequent. The antecedent and consequent of an association rule are disjoint: they do not have any items in common. To express the uncertainty in association rules, two measures are used: support and confidence. Support represents the number of transactions that include all items in the antecedent and consequent, and confidence is the ratio between the number of transactions that include all items in the consequent, as well as in the antecedent (namely, the support) and the number of transactions that include all items in the antecedent. A set of association rules for the purpose of classification is called class association rule set. A class association rule set is a subset of association rules with the specified classes as their consequents.

Over the course of last decade, the database community investigated the problem of rule mining with the specified classes as their consequences extensively, under the name of class or predictive association rule mining (these rules have the form: $\{A_1, A_2, \ldots, A_n \rightarrow Class\}$). The aim here is focused on using exhaustive search techniques to find all rules with the specified classes as their consequences that satisfy various interesting measures, such as minimum support and minimum confidence. Although class association rules can be discovered to a certain extent, they suffer from some drawbacks inherited from association rule mining. Firstly, both traditional and class association rule mining uses minimum support as an interestingness measure in the frequent itemset generation phase, which is inadequate for unbalanced class distribution: if the minimum support is high, class association miming will not generate sufficient rules for infrequent classes, while if the minimum support is too low, class association mining will generate over-fitting rules for frequent classes. Secondly, a large number of association rules in the training dataset will lead to a combinatorial explosion in the class association mining algorithms, which in turn, will not be able to generate rules that are important for the purpose of classification.

In our medical context, class association rule mining algorithms can be used to discover top $K$ associations of the above mentioned form, where $\{A_1, A_2, \ldots, A_n\}$ would be features/phenotypes and *Class* would be the disorder. However, due to the above listed reasons, these are not able to deal with characteristic features as per our definition. In this paper, we propose a novel class association mining algorithm that exploits an established interestingness measure – confidence – to model the discriminative aspect of characteristic features in conjunction with a new measure for pruning and finding class-based frequent features, hence addressing the first requirement of the definition of characteristic features.

Experimental results show that, based on a voting strategy classification evaluation, our proposed approach achieves a 3–10% increased accuracy when compared to traditional class association rule mining (from 30.94% to 47.50% against 27.04–37.24%), both subject to the recall cut-off point. In fact, our approach is able to discover more accurate characteristic features with an accuracy growth of 27.55%, a precision growth of 63.64% and a recall growth of 27.68% at recall cut-off point 5. Human-based validation, on the other hand, shows a positive correlation between the features deemed to be discriminative in a pair-wise disorder context and the pair-wise sensitivity and specificity of that disorder.

## 2. Materials and methods

### 2.1. Data characteristics

As mentioned previously, we have used the ESDN patient repository within our experiments. This consists of more than 1200 patient cases collectively acquired and discussed. The ESDN case workflow comprises three major steps: (i) a patient case is uploaded and an initial diagnosis is set by the original clinician that referred the case – patient cases contain a free text clinical summary and associated X-rays; (ii) the panel of experts discusses the case until an agreement is reached; and (iii) the panel of experts recommends a diagnosis.

In ESDN, each patient case includes a free text description of the clinical features, the relevant family history and a set of radiographic (X-ray) images. The free text clinical summary comprises all observed and relevant phenotypes of the patient, which can usually be validated via the radiographic images. The ESDN experts use this information to discuss possible diagnoses, and once an agreement is reached, the case receives a final diagnosis and is closed. The approach described in this paper uses ESDN's unique source of data for training and testing purposes. More specifically, we extracted clinical features from 1281 patient clinical summaries and recorded the initial and final diagnoses.

Since ESDN clinical summaries are in a free text format, they pose obvious challenges when aiming for efficient and automated knowledge discovery. Using the NCBO Annotator [6] and the Human Phenotype Ontology (HPO) [7] as background knowledge, we have performed automated concept extraction from the free text and defined phenotype feature sets for all patient cases. These extracted feature sets have then been used as input for knowledge discovery process. In order to get a better understanding of the concept recognition process, we refer the reader to Jonquet et al. [6].

More concretely, we have performed two data preprocessing steps. Firstly, we extracted patient phenotypes by annotating the text with corresponding terms from the Human Phenotype Ontology (HPO). In recent years, phenotype ontologies have been seen as an invaluable source of information, which can enrich and advance evolutionary and genetic databases [8]. HPO is currently the most comprehensive source of such information, comprising more than 10,000 terms organised in a hierarchical structure based on the anatomical localisation of the abnormality. The actual annotation process was performed using the NCBO Annotator [6], an ontology-based web service for annotation of textual sources with biomedical concepts. The annotation of a clinical summary resulted in a set of HPO terms. These have then been manually validated by a bone dysplasia expert, which led to a 100% correctness of the data used as input in our algorithm. Furthermore, to increase the processing speed, we have transformed both the HPO concepts, as well as the bone dysplasia diagnoses into a symbolic vector. For example, *short stature* is mapped to `S1`, *cleft palate* to `S2`, *Achondroplasia* to `D1`, and so on. The symbolic