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

vector associated with each patient is used as input for the key feature mining process. Each symbolic vector is also labelled with a disorder (class).

The ESDN dataset features 114 different types of skeletal dysplasias. The result of the preprocessing phase enabled us to perform a quantitative and qualitative analysis of the data. Firstly, we have found that the patient cases are not evenly distributed, i.e., less than half of the disorders had more than two cases. This has serious implications on any data mining algorithm as the general tendency will be to give preference to those disorders (classes) that are better represented. Fig. 1 shows the relative distribution of disorders according to the number of cases. It can be observed that 70% of the bone dysplasias have a very small number of cases (i.e., 1–2 or 3–5) and only 4% of the disorders are very well represented, i.e., they have more than 50 cases.

Secondly, by looking at the coverage of the clinical features, we have found (as expected) that the data is sparse. The coverage of a single feature can be defined as the percentage of cases diagnosed with a particular dysplasia in which this phenotype is present. For example, *cystic hygroma* has coverage of 50% in case of *Achondrogenesis type 1A* because it appears in 2 of total 4 cases diagnosed with this disorder. As a remark, we excluded from our analysis all cases that had listed a single phenotype. The maximum coverage achieved was 50% (e.g., *subglottic stenosis*), while the minimum was 0.99% (e.g., *immunodeficiency*). The average coverage was 11.33%, with a median of 8% and a mode of 0.99%. This exhibits the high sparsity of the data.

In order to achieve realistic results in our algorithm, from the 114 types of dysplasias present in the ESDN dataset, we chose only those that were represented by more than 10 patient cases. This has reduced our dataset to 394 annotated patient cases diagnosed with 15 different bone dysplasias (i.e., around 33% of the total number of cases). These cases are characterised by a total of 441 distinct phenotypes, with an average of 63.67 distinct phenotypes per disorder and an average of 4.49 distinct phenotypes per case. The final list of disorders and their associated number of cases is listed in Table 1.

2.2. Methodology

The goal of our work is to discover a set of key features given a disorder. More concretely, given a background knowledge base (i.e., annotated patient dataset) and a disorder, we aim to predict the top K characteristic features, ranked according to their probability. This is a classical problem of reasoning on disorder-feature

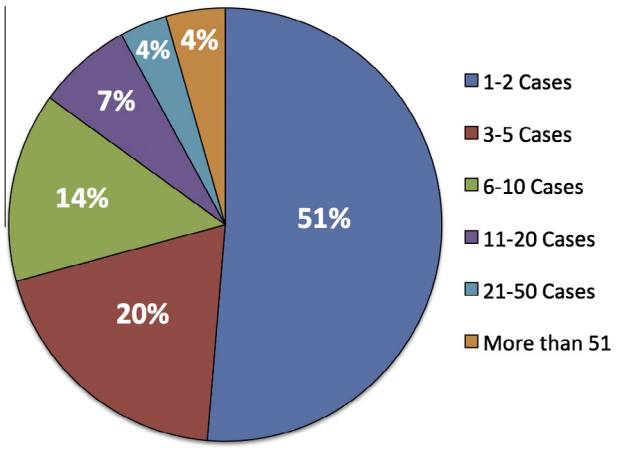

Fig. 1. Relative distribution of bone dysplasias according to the number of cases. More than 84% of the bone dysplasias present in the ESDN dataset have a very small number of cases (up to 10), while those that are well represented (i.e., over 50 cases) represent a mere fraction of the total number – 4%.

Table 1
Distribution of cases per disorder in the dataset used within our experiments.

Bone dysplasia	Number of cases
Achondroplasia	14
Cartilage-Hair-Hypoplasia	28
Cleidocranial dysplasia	10
Diastrophic dysplasia	11
Hypochondroplasia	22
Kniest dysplasia	14
Metaphyseal dysplasia, Schmid type	17
Multiple epiphyseal dysplasia	88
Osteogenesis Imperfecta	16
Osteopetrosis	11
Pseudoachondroplasia	44
rMED	20
Spondyloepiphyseal dysplasia congenital	71
Stickler syndrome	12
Thanatophoric dysplasia	16

associations in which class association rule mining algorithms are employed. However, since these rely on computing the frequency in the entire knowledge base, instead of using a class (disorder)-oriented fragmentation, they are unable to satisfy the first condition specified in our definition for characteristic features, and hence also the joint presence of both conditions. In order to comply with our definition, frequency needs to be considered in the context of a single disorder, i.e., frequency at class/disorder level. We, hence, propose a novel characteristic features ranking algorithm that uses a level wise search method to discover the interesting features.

The exact workflow of our method is listed below: (i) from a given set of free text clinical summaries, we extract a set of phenotypes corresponding to each patient case; (ii) these are then re-grouped according to the underlying diagnosis (disorder) and used in the process of discovering characteristic features; and (iii) finally, the resulting features are evaluated against the results produced by a typical class association mining algorithm.

Given a set of m patient cases: $\{P_1, P_2, \ldots, P_m\}$, each patient case consisting of n phenotypes and a diagnosed disorder D, our algorithm uses a scoring function $S(i)$ to compute the probability of each phenotype to represent a characteristic feature of the underlying disorder D. By convention, we assume higher probabilities/scores to correspond to more valuable characteristic features. As per the definition introduced earlier, the scoring function $S(i)$ takes into account, at the same time, two perspectives for each phenotype: *frequency* in the context of disorder D and *discriminative power* in the context of the other closely-related disorders.

We employ a measure named *Commonality* to take into account the frequency of the feature sets at class/disorder level, rather than the widely used support, which takes into account the frequency of the feature sets at the dataset level. The *commonality* of a phenotype set in the context of a particular disorder (class) represents the number of cases that include all phenotypes in the phenotype set and the disorder among the total number of phenotypes associated with that disorder. In other words, the *commonality* of a pattern for a class is the ratio between the number of transactions that include all features in the pattern and the class and the total number of transactions of that class. If there are n disorder classes $\{D_1, D_2, \ldots, D_n\}$ and p is a set of phenotypes, *commonality* is defined using the equation below.

$$Commonality(p|D_i) = \frac{\text{Total number of cases featuring } p \text{ in } D_i}{\text{Total number of cases of } D_i} \quad (1)$$

This measure is very intuitive when it comes to discovering characteristic features and it can easily be controlled by experts,

who can define how frequent should a phenotype set be in the context of any given disorder. On the other hand, this measure cannot be computed via standard class association mining algorithms, as they perform support calculation at the dataset level.

To define the second condition of our definition, i.e., *discriminative power*, we adapted the *confidence* measure. The *confidence* of a phenotype set in the context of a particular disorder represents the number of cases that include all phenotypes in the phenotype set and the disorder in the entire dataset. In other words, the *confidence* of a feature set for a class is the ratio between the number of transactions that include the feature set and the class and the total number of transactions that include the feature set in the entire dataset. If there are n disorder classes $\{D_1, D_2, \ldots, D_n\}$ and p is a set of phenotypes, *confidence* is defined using the equation below.

$$Confidence(p|D_i) = \frac{\text{Total number of cases featuring } p \text{ in } D_i}{\text{Total number of cases featuring } p} \quad (2)$$

The following section describes the novel algorithm we have designed to mine characteristic phenotypes from the ISDS dataset using the above defined measures.

2.3. The CFML algorithm

In order to discover characteristic phenotypes (class association rules), CFML (Characteristic Feature Mining algorithm) mines the training data by performing a class-wise grouping of the phenotypes that have the *confidence* over a certain threshold. Since mining is done class-wise, the algorithm will not be influenced by the uneven class distribution of the dataset. Furthermore, this makes it highly scalable and efficient, as the candidate generation phase can be massively parallelized.

Generally, all the algorithms interested in discovering feature sets in the data make multiple passes over the training set. In our case, in the first pass, we count the commonality measure of individual phenotypes in each class and determine which of them are frequent, i.e., have a minimum commonality. In each subsequent pass of a class, we start with a seed set of phenotype sets (previously found as frequent), generate new potentially frequent phenotype sets (i.e., candidate phenotype sets) and compute their commonality. At the end of the pass, we determine which of the candidate phenotype sets are actually frequent for each class, and use these as seeds for the next pass. The process stops when no new frequent phenotype sets are found.

The CFML algorithm is listed in Algorithm 1. The symbols used in the algorithm are defined below:

- n – number of classes
- K – level number
- G – number of all classes
- g_i – class i (a particular class)
- I_{GK} – desired feature sets at level K for all classes G
- I_G – desired feature sets for all levels and all classes G
- C_{GK} – candidate feature sets at level K for all classes G
- $C_{gi,k}$ – candidate feature sets at level K for the class g_i
- T_G – characteristic feature sets for all classes

Algorithm 1. Discovery of characteristic features

Require: $DataRows, min\_commonality, topK$
Ensure: Top ranked characteristic features for each class.
$K = 1$
I_{GK} = Select all 1-feature_sets with commonality greater or equal to *min_commonality*
while $I_k \neq \phi$ **do**
K++
for all $i \in 1 : n$ **do**
$C_{gi,k}$ = Candidate_generation ($I_{gi,k-1}$)
end for
Calculate Class-wise Count ($C_{GK}, DataRows$)
I_{GK} = Frequent Feature Set Selection ($C_{GK}, min.commonality$)
$I_G = I_G \cup I_{GK}$
end while
I_G = Measure Confidence Of Frequent Feature Sets (I_G)
I_G = Rank Features For Each Class (I_G)
T_G = Select TopK Feature Sets From Each Class ($I_G, topK$)
return T_G

Algorithm 2. Candidate_generation

Require: I_{k-1}
Ensure: Candidates.
for all $i \in I_{k-1}$ **do**
for all $j \in I_{k-1}$ **do**
$CF = i \cup j$
if $size(CF) == k$ **then**
Add to C_k if every subset of CF is frequent.
end if
end for
end for
return C_k

Algorithm 3. Calculate Class-wise Count

Require: $C_{GK}, DataRows$
Ensure: Count.
for all transaction t in $DataRows$ **do**
g_t = find the class of t
C_t = Find candidates of $C_{go,k}$ from the subsets of t
for all $c \in C_t$ **do**
$Count(c)$++
end for
end for
return $Count$

The first pass of the algorithm counts feature occurrences in each class to determine the frequent 1-featureset of each class. A subsequent pass, say pass k, consists of two phases. Firstly, the frequent feature sets $I_{gi,k-1}$, mined in the $k-1$th pass for class g, are used to generate the candidate feature set $C_{g,k}$, using the Candidate_generation function (Algorithm 2). This procedure is applied for every given class. Secondly, the dataset is scanned and the presence of candidates of each class is counted (Algorithm 3). The commonality of each candidate is calculated and the frequent feature set is constructed based on those candidates that are over a certain commonality threshold (using the FrequentFeatureSetSelection method – Algorithm 4). This process continues until there are no large frequent feature sets left. Once all the frequent feature sets are discovered, we compute the confidence value of each feature set (Algorithm 5), we rank the feature sets, sort them in a descendant order and finally return the top K items.

The actual ranking of the features in each class is computed in the following manner: given two frequent feature sets, F_a and F_b of a class C, F_a precedes F_b if:

1. the confidence of F_a is greater than that of F_b
2. the confidence values of F_a and F_b are the same, but the commonality of F_a is greater than that of F_b
3. the commonality and confidence values of F_a and F_b are the same, but $RC(F_a, C) > RC(F_a, C)$, where $RC(F, C)$ is the number of records of class C that match the conditions of X
4. the confidence, commonality and RC values of F_a and F_b are the same, but F_a has fewer conditions in its left hand side than of F_b
5. all criteria above are identical for F_a and F_b, but F_a was generated before F_b.

A novelty of our proposed algorithm is that it generates the candidate feature sets of a class in a pass by using only the feature sets of the particular class found frequent in the previous pass, without considering the feature sets of other classes in the dataset. The basic intuition is that any subset of a frequent feature set of a class must be frequent in that class. Therefore, the candidate feature set of a class having k features can be generated by joining frequent feature sets of the same class having $k-1$ features. This procedure results in the generation of a very small number of candidate feature sets.

Algorithm 4. Frequent Feature Set Selection

Require: $C_{GK}, DataRows$
Ensure: Frequent feature sets.
  $I_G = \{\}$
  for all $i \in 1 : n$ **do**
    for all $c \in C_{gi,k}$ **do**
      $Commonality(c) = \frac{Count(c) \in g_i}{\text{Total number of transactions in} g_i}$
      if $Commonality(c) >= min\_commonality$ **then**
        Add to c to I_G
      end if
    end for
  end for
  return I_G

Algorithm 5. Measure Confidence Of Frequent Feature Sets

Require: I_G
Ensure: Confidence of Frequent feature sets.
  $I = \{\}$
  for all $k \in 1 : size(I_G)$ **do**
    for all $i \in 1 : n$ **do**
      for all $f \in I_{gi,k}$ **do**
        $Confidence(f) = \frac{Count(f) \in g_i}{Count(f) \text{ in all groups}}$
      end for
    end for
  end for
  return I

2.4. Experimental setting

The quality of discovered characteristic features depends on their ability to determine the correct diagnosis. To measure accuracy, we have employed a weighted voting strategy [9], which allows all firing characteristic feature sets to contribute to the final prediction and is described below in more detail.

A voting-based classification method classifies a new instance according to the number of characteristic feature sets covering it. Voting allows all firing characteristic feature sets to contribute to the final prediction. This strategy combines the characteristic feature sets $CF(x)$ that fire upon a new patient case x. A simple voting strategy considers all the rules in $CF(x)$. Given D a set of n disorders $\{d_1, d_2, \ldots, d_n\}$, we denote the class voted by a characteristic feature set k with a binary function $vote(k, d_i)$ that takes the value 1 when k votes are received for disorder d_i, and 0 for the any other class. The disorder that receives the maximum number of votes is the most probable diagnosis for case x.

$$TotalVote(d_i) = \sum_{k \in CF(X)} Vote(k, d_i) \tag{3}$$

Weighted voting is similar to simple voting, however, each vote is multiplied by a factor that quantifies the quality of the vote.

$$TotalVote(d_i) = \sum_{k \in CF(X)} Vote(k, d_i) * q(k, d_i) \tag{4}$$

where $q(k, d_i)$ is the quality of vote.

To assess the efficiency of our proposed algorithm against a standard class association rule mining algorithm (CARM), we have calculated the overall accuracy of the discovered characteristic features. In all experiments, we compute the prediction accuracy as the overall percentage of correctly predicted disorders at a given recall cut-off point (i.e., by taking into account only the top K predictions for different values of K, where K is the recall cut-off point). Hence, a success represents a correctly predicted disorder (the exact same, and not a sub or super class of it), while a miss represents an incorrectly predicted disorder. If N is the total number of test cases and L is the number of correctly predicted disorders, then accuracy $A = L/N$. This is expressed in percentages in Table 2 and in the Results section. CARM has been implemented as an adapted Apriori algorithm [10].

In order to provide an accurate view over the prediction of the discovered key features, each experiment has been performed as a 5-fold cross validation with an 80–20 split (80% knowledge base, 20% test data). Table 2 lists the resulted average accuracy at five different recall cut-off points. Finally, we have used a maximum size for the characteristic feature sets of 10 as the computational cost increases exponentially with the feature set size and item set size in both our proposed algorithm, as well as in the standard class association rule mining process. The goal of our approach is to discover the top K characteristic features, and within our experiments K has been set to 5.

A final remark should be made with respect to the different parameters that can be tuned in both algorithms. To ensure a fair and correct evaluation, we performed experiments to determine

Table 2
Experimental results: CFML vs. traditional class association rule mining (CARM).

Algorithm	Accuracy@1 (%)	Accuracy@2 (%)	Accuracy@3 (%)	Accuracy@4 (%)	Accuracy@5 (%)
CFML	30.95	40.33	44.74	46.68	47.50
CARM	27.04	33.39	35.86	37.24	37.24
Growth (%)	14.42	20.78	24.76	25.34	27.55

the optimal parameter values in each case, i.e., the set of parameters that discover the highest quality characteristic features. These parameters have subsequently been used within our experiments. For our approach, we had to determine the best *commonality* value, and as shown in Fig. 2, a value of 14% provides the best characteristic features. Similarly, for the CARM approach, we had to determine the optimal *support* value, where support is defined as the coverage of the feature sets in the total number of cases. As shown in Fig. 3, the best *support* value is 1.07% (i.e., $4/374 * 100 - 4$ feature sets, 374 total number of cases). For clarification purposes, the best set of characteristic features is defined as the set that provides the best classification accuracy.

3. Results and discussion

In this section we discuss the experimental results achieved by applying both CFML and CARM on the ESDN dataset. We start by looking at the classification results via macro accuracy and class-based precision and recall, then we discuss the pair-wise discriminative power of our algorithm and finally, we perform a human-based evaluation of the resulting characteristic features.

3.1. Classification results

As previously mentioned, class association rule mining uses rules of the form $\{S \rightarrow D\}$ to discover characteristic features, where S is a set of features/phenotypes and D is the set of disorders. In order to rank and compare the resulting features we have applied the same methodology to both CFML (our algorithm) and CARM (an implementation of standard class association rule mining) – i.e., by using *confidence* as a primary measure for ranking purposes.

Table 3
Experimental results: overall comparative precision and recall across the both approaches.

	CFML		CARM	
	P (%)	*R* (%)	*P* (%)	*R* (%)
Achondroplasia	0.00	0.00	6.67	40.00
Cartilage-Hair-Hypoplasia	44.67	46.67	6.50	36.67
Cleidocranial dysplasia	80.00	60.00	0.00	0.00
Diastrophic dysplasia	36.67	60.00	45.00	60.00
Hypochondroplasia	29.33	30.00	0.00	0.00
Kniest dysplasia	30.00	20.00	29.00	30.00
Metaphyseal dysplasia, Schmid type	25.00	20.00	28.33	30.00
Multiple epiphyseal dysplasia	26.40	16.98	0.00	0.00
Osteogenesis Imperfecta	0.00	0.00	0.00	0.00
Osteopetrosis	20.00	20.00	0.00	0.00
Pseudoachondroplasia	25.00	19.50	32.86	27.74
rMED	24.00	18.33	16.67	18.33
Spondyloepiphyseal dysplasia congenital	26.94	18.69	20.00	7.27
Stickler syndrome	20.00	10.00	40.00	30.00
Thanatophoric dysplasia	13.00	30.00	20.00	10.00
Average	26.73	24.68	16.33	19.33
Average precision growth	63.64%			
Average recall growth	27.68%			

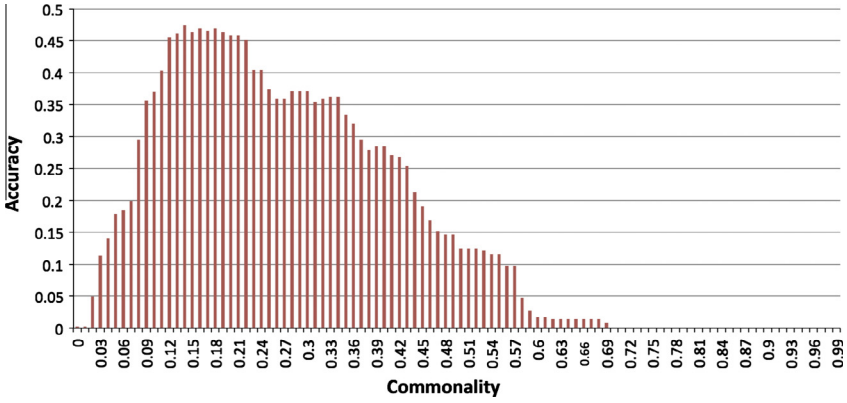

Fig. 2. Experimental results to determine the optimum *commonality* value for our approach. The optimal commonality value is achieved by considering the commonality value across all disorders for which the overall accuracy is maximised.

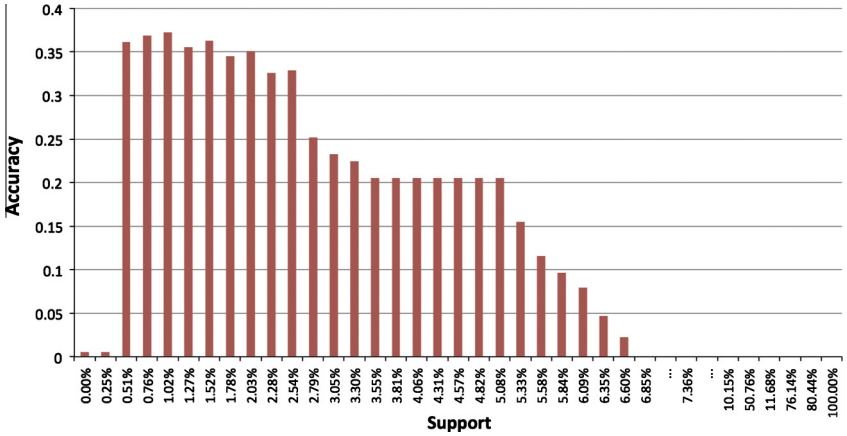

Fig. 3. Experimental results to determine the optimum *support* value for standard class association rule mining algorithm. The optimal support value is computed by looking at the number of cases that list a particular phenotype normalised by the total number of cases in the dataset for which the accuracy is optimal.

The *support* of the characteristic features within the disorder set *D* is one of the most important contributing factors in the prediction accuracy, and it is proportional to it. To take *support* into account, both CARM and CFML use a minimum threshold – CARM for standard support and CFML for commonality (i.e., the class-oriented support). The other contributing factor is the rarity of the characteristic features, which is represented by *confidence* in both CARM and CFML.

Using the optimal commonality and support described in the previous section (see Figs. 2 and 3), we have performed a fivefold cross validation with stratification and computed the average macro accuracy of the two methods at five different recall cut-off points. Table 2 lists the results. Overall, it is clear that the sparsity of the data has heavily affected both approaches, as the final accuracies are fairly low (this aspect is also shown in the phenotype distribution discussed in the Data characteristics section). Nevertheless, CFML has outperformed CARM, hence showing that it is able to discover patterns specific to particular classes/disorders, instead of patterns frequent in the entire dataset. Furthermore, this result is achieved in the context of an unbalanced distribution of classes in the ESDN dataset, which causes more issues to CARM rather than our approach. Finally, we can observe a bigger growth in accuracy with the increase in the recall cut-off point, from 14.42% at $K = 1$ to 27.55% at $K = 5$.

The feature sets of each disorder in our dataset have different underlying characteristics, and hence it is expected to achieve different results for different classes, with the relative distribution of the features in the set of cases being an influencing factor. In order to understand the efficiency of the two methods at a lower level, we have computed the class-based precision and recall, as listed in Table 3. Overall, our approach outperforms CARM with an average precision growth of 63.64% and an average recall growth of 27.68%. CFML achieves a 26.73% average precision and 24.68% average recall, whereas CARM achieves a 16.33% average precision and 19.33% average recall.

There results reveal a series of aspects that are worth noting. Firstly, CARM is unable to get any result in the case of four disorders where CFML scores reasonably high. Secondly, the situation is being inverted in the case of one disorder – i.e., Achondroplasia. Unlike CARM, CFML has performed fairly uniform across all classes

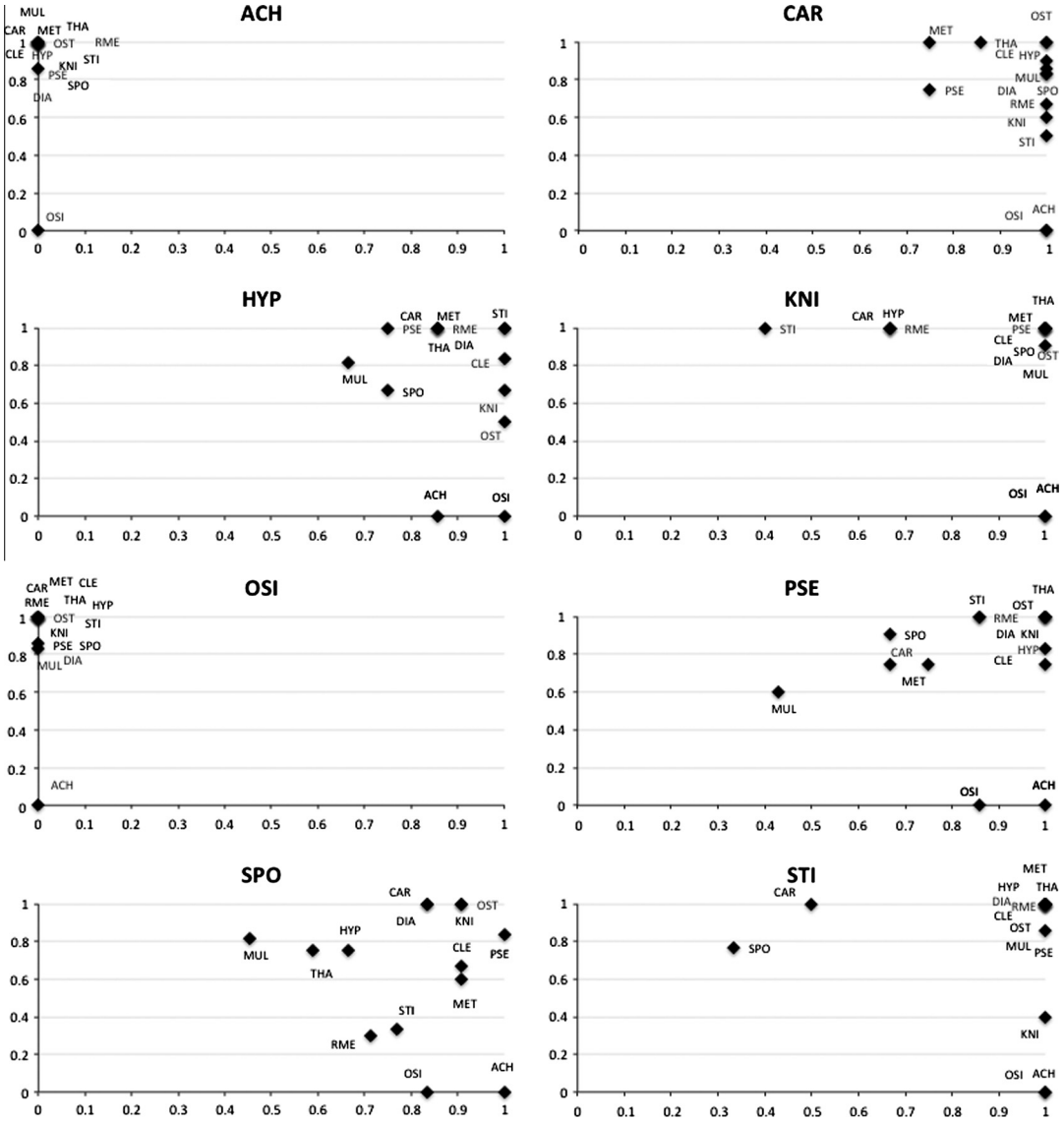

Fig. 4. CFML pair-wise specificity–sensitivity analysis. Each graph depicts the specificity and sensitivity of the disorder under scrutiny against all other (the *X*-axis denotes specificity and the *Y*-axis denotes sensitivity). The higher the values, in particular the clusters in the upper-right corner, the more discriminative the current disorder is. The characteristic features chosen by CFML were highly or sufficiently discriminative in 12 our of the total 15 disorders.

because it treats all classes equally, independently of their frequency. Infrequent disorders, such as Cleidocranial dysplasia and Osteopetrosis, have only 10–11 cases in the dataset, while the average number of cases per disorder is 26.67. Hence, they are particularly problematic for the dataset-oriented approach followed by CARM. For example, CARM discovered *short stature*, *sparse hair* or *frontal bossing* as being characteristic for Cleidocranial dysplasia, however they did not have enough discriminative power, as seen also from their *confidence* values 0.1, 0.09 and 0.04 respectively. On the other hand, in the case of Osteopetrosis CARM was not able to mine any features because the set of phenotypes corresponding to this disorder are not sufficiently frequent when compared to the overall set of phenotypes in the dataset.

If the average phenotype coverage of a particular disorder is lower than the overall average coverage of phenotypes, CARM tends to select those features present in many other disorders, whereas CFML treats the phenotype set of each disorder independently. For example, the average phenotype coverage of Hypochondroplasia and Multiple epiphyseal dysplasia is 7.4% and 2.7% respectively and the overall mean coverage is 9.35%. CARM discovered *short stature* to be characteristic for Multiple epiphyseal dysplasia and eight other disorders, which clearly shows its low discriminative power. In the case of Hypochondroplasia, CFML selected *relative macrocephaly* (0.27) and *short palm* (0.19) – characteristic features of this single disorder – and achieved much better results.

As shown, Achondroplasia was problematic for CFML, although the selected characteristic features (e.g., *lumbar kyphosis*, *trident abnormality* or *low nasal bridge*) had very high *confidence* values (1.0, 1.0 and 1.0). Here, the *commonality* of the features – on average 0.17 in all Achondroplasia patient cases – raised issues because it was too low. CARM selected *macrocephaly*, *rhizomelic shortening* and *frontal bossing*, which did not have high *confidence* values (0.07, 0.06 and 0.05) but were very frequent in the entire dataset – with an average *support* of 0.45.

Osteogenesis imperfecta proved to be an issue for both approaches since none of them was able to correctly classify corresponding cases. CARM selected a very generic feature to be characteristic – *skeletal dysplasia* – with an extremely low *confidence* value of 0.04 – not enough to identify the underlying

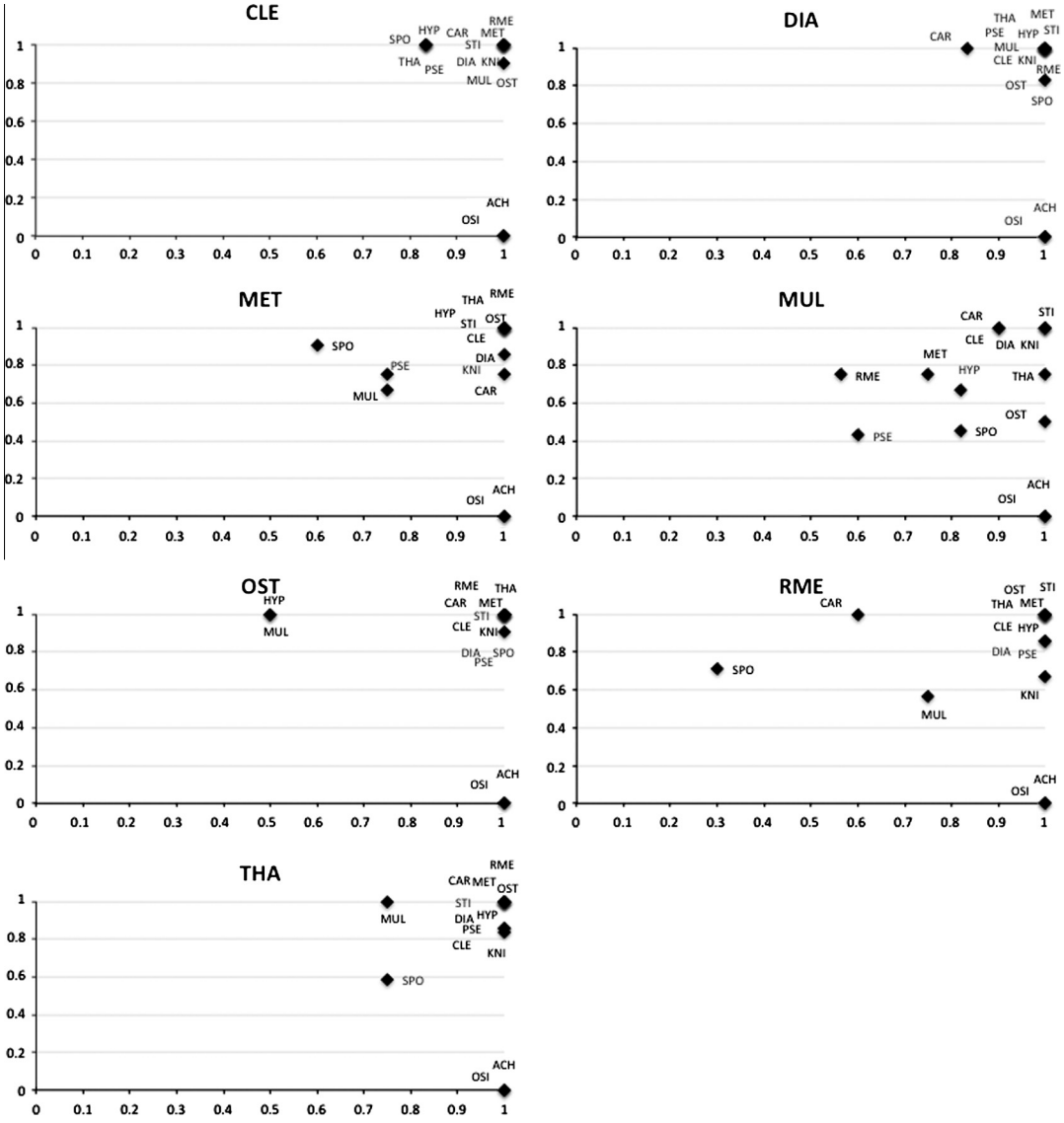

Fig. 5. CFML pair-wise specificity–sensitivity analysis. Each graph depicts the specificity and sensitivity of the disorder under scrutiny against all other (the *X*-axis denotes specificity and the *Y*-axis denotes sensitivity). The higher the values, in particular the clusters in the upper-right corner, the more discriminative the current disorder is. The characteristic features chosen by CFML were highly or sufficiently discriminative in 12 our of the total 15 disorders.

Table 4
Legend of disorder symbols used in the experimental results.

Symbol	Disorder
ACH	Achondroplasia
CAR	Cartilage-Hair-Hypoplasia
CLE	Cleidocranial dysplasia
DIA	Diastrophic dysplasia
HYP	Hypochondroplasia
KNI	Kniest dysplasia
MET	Metaphyseal dysplasia, Schmid type
MUL	Multiple epiphyseal dysplasia
OSI	Osteogenesis Imperfecta
OST	Osteopetrosis
PSE	Pseudoachondroplasia
RME	rMED
SPO	Spondyloepiphyseal dysplasia congenital
STI	Stickler syndrome
THA	Thanatophoric dysplasia

disorder. CFML, on the other hand, discovered *recurrent fractures*, *short femur* or *short lower limbs* as features, with high *confidence* values – 1.0, 0.5 and 0.67 respectively – but with low commonality (18.75%), which led to the patient cases being classified in different other classes.

3.2. Pair-wise sensitivity and specificity

As we have seen in the previous section, from a macro perspective, our approach performs uniformly across all disorders (with a few exceptions). In order to get a better insight into the individual discriminative power of the characteristic features selected by CFML, we have compiled pair-wise confusion matrices for each disorder and computed the pair-wise sensitivity and specificity. Figs. 4 and 5 depict 15 sensitivity–specificity graphs corresponding to the 15 disorders in our dataset (the legend of disorder symbols used by the figure and the description below can be found in Table 4 – also the *X*-axis denotes specificity and the *Y*-axis denotes sensitivity).

In general, sensitivity represents the probability of a positive outcome given an underlying positive element (e.g., probability of a positive test, given that the patient is ill), while specificity represents the probability of a negative outcome given an underlying positive element (e.g., probability of a negative test, given that the patient is well). In our context, sensitivity denotes the ability of the chosen feature set to correctly identify the disorder, while specificity denotes the ability to correctly identify that the real disorder is not the one from which the feature set has been selected.

Using the characteristic features selected by CFML we can observe that, in general, most disorders form specific discriminative pairs that could be applied in a differential diagnosis setting. For example, MUL is highly sensitive and specific to STI, KNI, DIA and CLE, PSE is sensitive and specific to THA, OST, STI and RME, or SPO is sensitive and specific to KNI and OST. However, the highlights are provided by two disorders (CLE and DIA) that have a uniformly high sensitivity and specificity against more than 85% of the other disorders (the problematic OSI and ACH are the only ones missing). This implies that the characteristic features inferred by CFML are particularly selective and describe very well these disorders. Similarly, the features of four other disorders (KNI, OST, THA and STI) performed fairly well and led to high sensitivity and specificity against 65% of the other disorders.

3.3. Human-based validation

In order to validate our results from a real-world perspective (i.e., to test to some extent the clinical significance of the resulted characteristic features), we have also performed a human-based

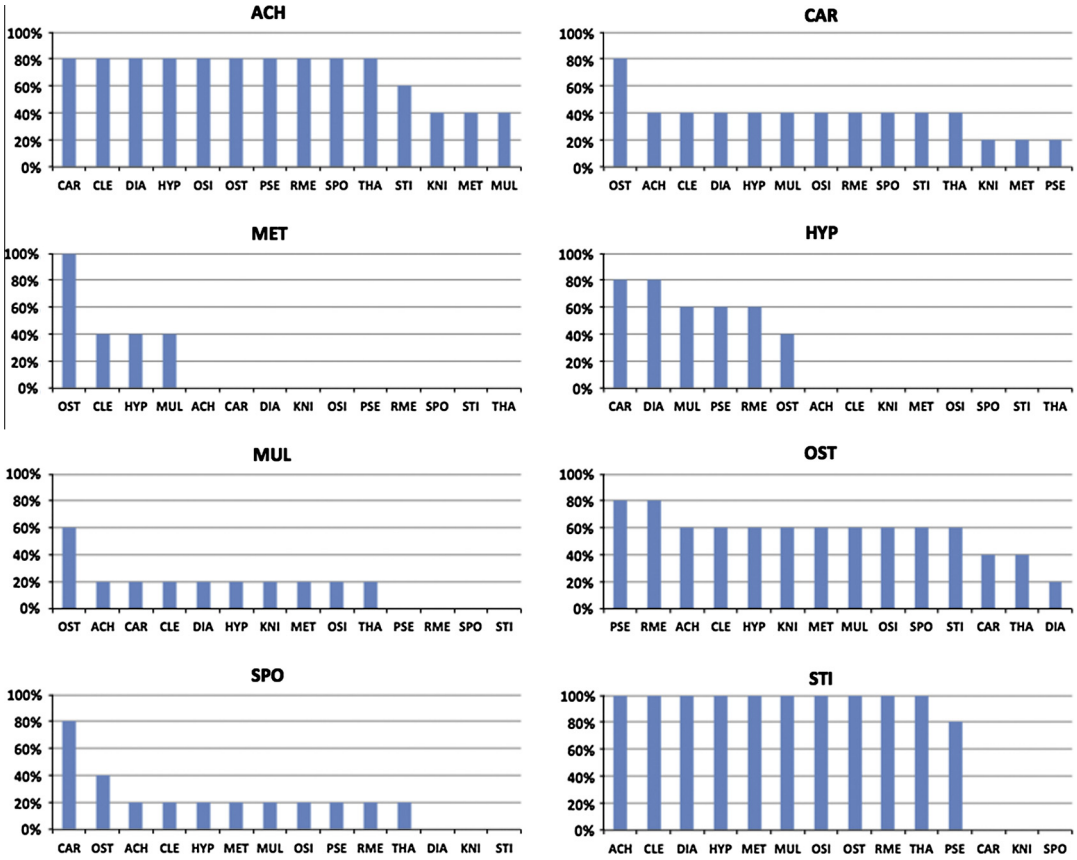

Fig. 6. CFML pair-wise discriminative power as established by an expert. In average 41.88% of the key features of all disorder are discriminative, while 58.12% are not discriminative.

validation. The experiment was carried out with the help of a bone dysplasia expert and comprised two parts. Firstly, the expert analysed, individually the set of characteristic features corresponding to each disorder, in order to determine their meaningfulness. More concretely, the domain expert went through all characteristic features inferred for every disorder and marked them as meaningful, possibly meaningful or not meaningful. We have then compiled statistics from this discrete categorisation. Here, the goal was to validate the selected features by judging if they are, in reality, phenotypes associated with the particular disorder, or just observations noted part of the clinical summary.

Secondly, the expert performed a pair-wise discriminative analysis and judged to what extent the characteristic features of a disorder are discriminative against a second disorder. To be more precise, the domain expert marked every characteristic feature of every disorder true or false to indicate whether it is discriminative against all the other disorders, in a pairwise manner. For instance, to note that *Lumbar kyphosis* is a characteristic feature of Achondroplasia that is discriminative against Cartilage-Hair-Hypoplasia, the expert marked it has true in the pairwise context Achondroplasia – Cartilage-Hair-Hypoplasia. Following this process, we have compiled the average discriminative values of the characteristics features of all disorders against all other disorders – these are depicted in Figs. 6 and 7.

The first part of the experiment revealed very good results. In average, 79.56% of the selected features were deemed meaningful, 17.78% were considered possibly meaningful, subject to a given context, while only 2.66% of the features were not meaningful. More concretely, the disorders that achieved high sensitivity and specificity scores in the previous experiment also had meaningful features, e.g., CLE – 80%, DIA – 100%, THA – 80%, STI – 100%, OST – 100%. Furthermore, not surprisingly, the problematic disorders, such as OSI achieved low meaningfulness scores – 25%.

The results of the second part of the experiment are depicted in Figs. 6 and 7, where each graph shows the percentage of features deemed to be discriminative in the context of a particular disorder against the others (the legend of disorder symbols can be found in Table 4). For example, in around 71% of the cases the features selected for STI were 100% discriminative, while in one case (against PSE) 80% of the features were discriminative and in the rest (21%) no features were discriminative. Overall, we can observe that 50% of the disorders achieved uniformly good results, the highlights being provided by STI, ACH, DIA, CLE and THA. In order to validate the results of the automatic classification, we tried to find correlations (the Pearson correlation coefficient) between the pair-wise sensitivity and specificity of the disorders and the percentage of discriminative features as indicated by the expert. In the case of sensitivity (i.e., the ability to detect true positives), positive correlations have been found for CLE (0.42), CAR (0.50), MUL (0.51), PSE (0.31) and STI (0.76), which reinforces the results presented in the previous experiment, especially for CLE and STI which have been found highly sensitive to most of the other disorders.

Similarly, in the case of specificity (i.e., the ability to detect true negatives), we found positive correlations for CLE (0.36), HYP (0.41), PSE (0.25), RME (0.27) and THA (0.46).

3.4. Related work

In biomedical domain, many researchers used class association rule mining or predictive association rule mining to solve classification problems or to discover various patterns in medical data [11–15]. All existing previous work relies on a direct application of the standard association rule mining algorithms, such as Apriori [10], FP-growth [5] or Eclat [5]. Typically, classes (disorders) of interest are specified and targeted as consequents of the association rules. However, these standard algorithms have been designed

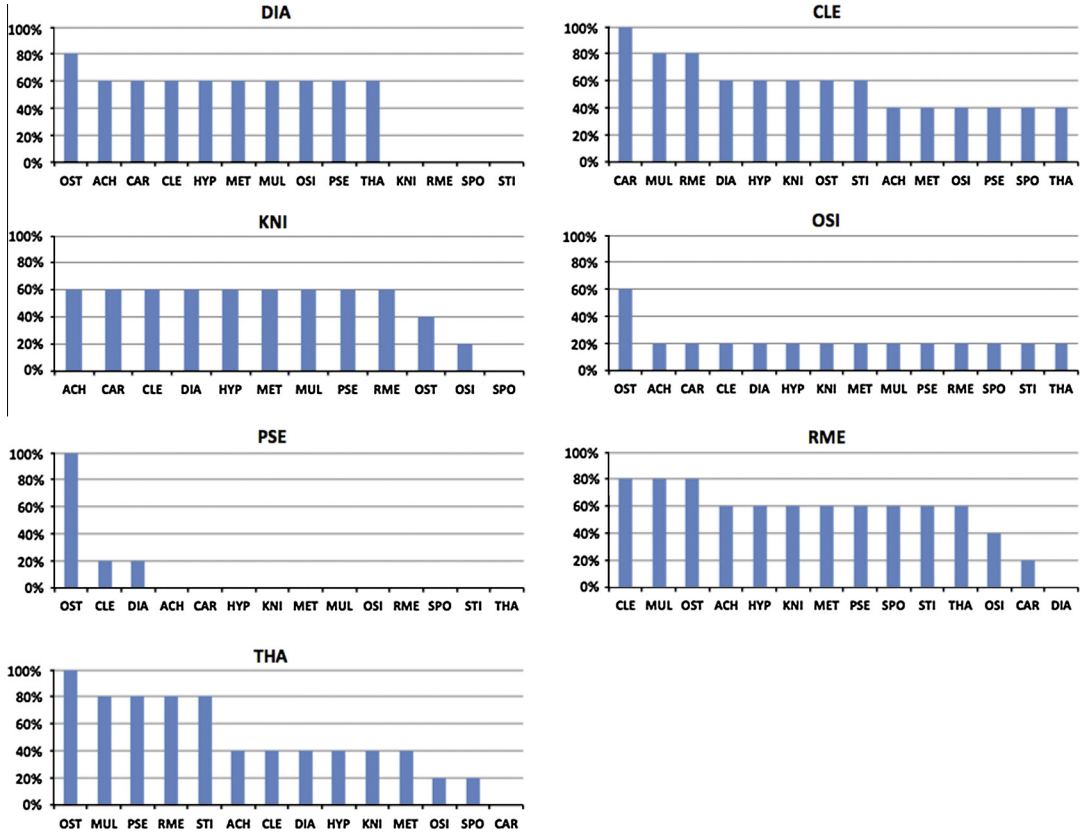

Fig. 7. CFML pair-wise discriminative power as established by an expert. In average 41.88% of the key features of all disorder are discriminative, while 58.12% are not discriminative.

for unlabelled data, hence they function in an unsupervised manner, while class association mining tasks takes advantage of the fact that the data is labelled, thus making the task supervised. Furthermore, these two types of approaches also differ in the way in which they use the data – standard algorithms find rules based on features frequently present in the entire dataset, while class association mining requires class-specific frequency.

Specific examples of approaches similar to ours, but suffering from the issues mentioned above include: (i) the work of Osl et al. [16] on identifying biomarker candidates in prostate cancer data or of Karabatak et al. [15] on detection of breast cancer – Apriori has been adapted to find all class association rules with support and confidence greater than some given thresholds; (ii) extracting meaningful patterns in Oriental Medicine [17], where Apriori has been adapted to discover class association rules by considering symptoms antecedent feature sets, and the herbal materials as consequent feature sets – classic support and confidence have been again the main underlying measures used; or (iii) predicting protein–protein interactions (PPI) [18] by generating class association rules where the consequent of the target class association is restricted to one of the PPI types in focus – the methodology and setting is similar to the ones above, i.e., adapted Apriori with support and confidence thresholds.

The nature of traditional association rule mining algorithms (i.e., their designated goal of working with unlabelled data) makes them inadequate for mining class association rules. Consequently, we have focused on devising an algorithm that takes advantage of this limitation, by class-specific generating candidates, and hence reducing the computation time significantly as well as improving the quality of the results. Furthermore, in the particular context set by the definition of characteristic phenotypes, our algorithm is able to mine features both frequent to a given class (instead of the entire dataset) and rare in the closely-related classes. In practice, this helps in discovering more specific rules for classes and to solve the unbalanced class distribution problem of existing class association algorithms.

4. Conclusion

In this manuscript we have presented a novel algorithm for mining characteristic phenotypes for skeletal dysplasias. We started by assigning a clear definition for *characteristic* phenotypes and then proposed a set of measures (*commonality* and *confidence*), together with an associated class-driven algorithm, to discover the top K such phenotypes in the context of a set of disorders and patient cases present in the ESDN repository. The experimental results show that, given a reasonable amount of data (considering the focus on rare diseases), our approach discovers more accurate characteristic features than a standard class association rule mining approach, achieving an accuracy growth of 27.55% at recall cut-off point 5.

Future research will focus on discovering multi-level characteristic features, which will require considering the Human Phenotype Ontology annotations used to model patient phenotypes at multiple levels of abstraction. For example, we will consider using a generalisation strategy for ontology traversal where the level of abstraction of the annotations is increased one level at a time on the each iteration of the characteristic feature mining process.

Acknowledgments

This research is funded by the Australian Research Council (ARC) under the Linkage grant SKELETOME – LP100100156 and the Discovery Early Career Researcher Award (DECRA) – DE120100508.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2013.12.001.

References

[1] Warman ML et al. Nosology and classification of genetic skeletal disorders: 2010 revision. Am J Med Genet A 2011;155:943–68.
[2] Li W, Han J, Pei J. CMAR: accurate and efficient classification based on multiple class-association rules. In: Proceedings of the 2001 international conference on data mining. San Jose, CA, US; 2001. p. 369–76.
[3] Jensen S. Mining medical data for predictive and sequential patterns. In: Proceedings of the PKDD discovery challenge on thrombosis data. Freiburg, Germany; 2001.
[4] Thabtah F, Cowling P, Peng Y. MCAR: multi-class classification based on association rule. In: Proceedings of the 3rd ACS/IEEE international conference on computer systems and applications. Cairo, Egypt; 2005. p. 33.
[5] Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. Data Min Knowl Discov 2007;15:55–86.
[6] Jonquet C, Shah NH, Musen MA. The open biomedical annotator. In: Proceedings of the 2010 AMIA summit of translational bioinformatics. San Francisco, CA, US; 2010. p. 56–60.
[7] Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. The Am J Hum Genet 2008;83:610–5.
[8] Mabee PM, Ashburner M, Cronk Q, Gkoutos GV, Haendel M, Segerdell E, et al. Phenotype ontologies: the bridge between genomics and evolution. Trends Ecol Evol 2007;22:345–50.
[9] Zhu X, Song Q, Jia Z. A weighted voting-based associative classification algorithm. Comput J 2010;53:786–801.
[10] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th international conference on very large data bases. Santiago de Chile, Chile; 1994. p. 487–99.
[11] Patil BM, Joshi RC, Toshniwal D. Association rule for classification of type-2 diabetic patients. In: Proceedings of the 2010 second international conference on machine learning and computing. Washington, DC, US; 2010. p. 330–4.
[12] Papageorgiou EI. A new methodology for decisions in medical informatics using fuzzy cognitive maps based on fuzzy rule-extraction techniques. Appl Soft Comput 2011;11:500–13.
[13] Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 2010;43:891–901.
[14] Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinform 2010;11:S7.
[15] Karabatak M, Ince MC. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl: An Int J 2009;36:3465–9.
[16] Osl M, Dreiseitl S, Pfeifer B, Weinberger K, Klocker H, Bartsch G, et al. A new rule-based algorithm for identifying metabolic markers in prostate cancer using tandem mass spectrometry. Bioinformatics 2008;24:2908–14.
[17] Yang DH, Kang JH, Park YB, Park YJ, Oh HS, Kim SB. Association rule mining and network analysis in oriental medicine. PLoS One 2013;8:e59241.
[18] Park SH, Reyes JA, Gilbert DR, Kim JW, Kim S. Prediction of protein–protein interaction types using association rule based classification. BMC Bioinform 2009;10.