Combining APRIORI and Bootstrap Techniques for Marker Analysis

Giacomo Gamberoni[1], Evelina Lamma[1], Fabrizio Riguzzi[1], Chiara Scapoli[2], and Sergio Storari[1]

[1] ENDIF, University of Ferrara, Italy
{giacomo.gamberoni, evelina.lamma, fabrizio.riguzzi, sergio.storari}@unife.it
[2] Department of Biology, University of Ferrara, Italy, scc@unife.it

Abstract. In genetic studies, complex diseases are often analyzed searching for marker patterns that play a significant role in the susceptibility to the disease. In this paper we consider a dataset regarding periodontitis, that includes the analysis of nine genetic markers for 148 individuals. We analyze these data by using a novel subgroup discovering algorithm, named APRIORI-B, that is based on APRIORI and bootstrap techniques. This algorithm can use different metrics for rule selection. Experiments conducted by using as rule metrics novelty and confirmation, confirmed some previous results published on periodontitis.

1 Introduction

In classical genetics [1], diseases are divided into Mendelian disorders and complex traits. While the former are attributed to single gene mutations with a simple mode of inheritance, the latter are thought to result from interaction among multiple genes. The main task in the study of these polygenic diseases is obviously to find the genetic patterns that increase susceptibility to the diseases.

In machine learning, such task is faced by using subgroup discovery techniques. Their goal is to find subgroups, represented by rules, which describe subsets of the population that are sufficiently large and statistically unusual with respect to a target attribute. This task is at the intersection of predictive and descriptive induction, and has been formulated in [2], [3], [4]. The problem can be expressed as follows: given a population and a single property of the individuals, find population subgroups that are statistically "most interesting". For example, we may look for groups that are as large as possible and on which the property of interest has a distribution that is as different as possible with respect to the distribution over the whole population. In the literature, several algorithms have been proposed for subgroup discovery (e.g. Explora [2], MIDOS [3], APRIORI-SD [5], CN2-SD [6]) and for classification rule learning (e.g. CBA [7]).

In this paper, we present a novel algorithm, named APRIORI-B, that performs subgroup discovery by combining APRIORI [8] and bootstrap techniques (more precisely the randomization test).

Our method uses APRIORI for finding frequent itemsets, and then generates rules from them. In the rule selection post-processing phase, it sorts the generated rules by using a rule evaluation metric. Then the most significant rules are selected by using the randomization test [9].

We verified the suitability of APRIORI-B for marker analysis by applying it on real biological data. In the experiment, we analyzed a dataset used by biologists to investigate the relation between nine genetic markers and periodontitis. For this biological dataset we provide some subjective evaluations of the subgroups identified.

This paper is organized as follows: Section 2 presents background information on APRIORI algorithm and methods for rule evaluation. Section 3 describes our algorithm. Section 4 illustrates the chosen case study: the analysis of genetic markers. Section 5 reports the results of applying our algorithm the genetic dataset. Finally, Section 6, presents conclusions and perspectives for future works.

2 Background

In subgroup discovery, subgroups can be modeled by classification rules. In this section, we first present association rules and then one of their special case, represented by classification rules (Section 2.1). Then in Section 2.2, we briefly describe the APRIORI algorithm [8] for association rule mining.

2.1 Association and classification rules

Association rules. Consider a table D having only discrete attributes. If D has also numeric attributes, they are discretized. An *item* is a literal of the form $A = v$ where A is an attribute of D and v is a value in the domain of A. Let M be the set of all the possible items. An *itemset* X is a set of items, i.e. it is such that $X \subseteq M$. A k-itemset is an itemset with k elements. We say that a record r of D *contains* an itemset X if $X \subseteq r$ or, alternatively, if r satisfies all the items in X. Let $n(X)$ be the number of records of D that contain X. Let $n(\overline{X})$ be the number of records of D that do not contain X. Let N be the number of records of D. The *support* of an itemset X (indicated by $Sup(X)$) is the fraction of records in D that contain X. i.e., $Sup(X) = n(X)/N$. It is also equal to the probability of a record of D of satisfying X, i.e. $p(X) = Sup(X)$. When X and Y are two itemsets we use the shorthand notation $n(XY)$, $Sup(XY)$ and $p(XY)$ to mean, respectively, $n(X \cup Y)$, $Sup(X \cup Y)$ and $p(X \cup Y)$.

Association rules are of the form $B \rightarrow H$ where B and H are itemsets such that $B \cap H = \emptyset$. B and H are respectively called *body* and *head*.

Classification rules. Classification rules are association rules whose head is of the form $Class = c$ where $Class$ is a special attribute of D. In this case, the records of D are also called *examples* and a rule $B \rightarrow Class = c$ covers a record r if $B \subseteq r$ and correctly covers a record if $B \cup \{Class = c\} \subseteq r$.

Notice that, for classification rules, a contingency table is a generalization of a confusion matrix, which is the standard basis for computing rule evaluation measures in binary classification problems. In the confusion matrix notation, $n(H)$ is the number of positive examples, $n(\overline{H})$ the number of negative examples, $n(B)$ is the number of examples covered by the rule therefore predicted as positive, $n(\overline{B})$ is the number of the examples not covered by the rule and therefore predicted as negative, $n(BH) = TP$ is the number of true positives, $n(\overline{BH}) = TN$ is the number of true negatives, $n(B\overline{H}) = FP$ is the number of false positives, and $n(\overline{B}H) = FN$ is the number of false negatives.

Rule metrics For association and classification rules a number of quality metrics can be defined. All rule evaluation measures are defined in terms of frequencies from the *contingency table* only (see Table 1).

Table 1. A contingency table.

	Body		
Head	B	$\overline{B}$	
H	$n(HB)$	$n(H\overline{B})$	$n(H)$
$\overline{H}$	$n(\overline{H}B)$	$n(\overline{H}\overline{B})$	$n(\overline{H})$
	$n(B)$	$n(\overline{B})$	N

Given a rule $R = B \rightarrow H$, we define the following metrics:

- Support: $Sup(R) = p(BH) = Sup(BH) = \frac{n(BH)}{N}$
- Confidence: $Conf(R) = p(H|B) = \frac{Sup(BH)}{Sup(B)} = \frac{n(BH)}{n(B)}$
- Novelty: $Nov(R) = p(HB) - p(H)p(B)$
- Confirmation: $Confirmation(R) = \frac{p(BH)-p(B)p(H)}{\sqrt{p(B)p(H)p(\overline{B})p(\overline{H})}}$

Support and Confidence are classical association and classification rule metrics. Novelty [10] and Confirmation [11] are examples of more complex rule evaluation metrics [12], and we choose to focus the experiments described in this paper on them.

The definition of novelty states that we are only interested in high support if that could not be expected from the marginal probabilities, i.e., when $p(H)$ and/or $p(B)$ are relatively low. It can be demonstrated that $-0.25 \leq Nov(R) \leq 0.25$: a strongly positive value indicates a strong association between H and B, while a strongly negative value indicates a strong association between $\overline{H}$ and B.

2.2 APRIORI

The task of discovering association rules consists in finding all the association rules having a minimum support *minsup* and a minimum confidence *minconf*.

In order to discover such rules, the approach proposed in [8] first discovers all the itemsets with support higher than *minsup* and then finds the rules from them. The itemset with support above *minsup* are called *large*. The part of APRIORI that finds large itemsets is shown in Figure 1. Figure 2 shows function apriori-gen that is used by APRIORI.

Notation:L_k, set of large k-itemset

1. $L_1 = \{$ large 1-itemsets $\}$
2. for(k=2; $L_{k-1} \neq \emptyset$; $k++$) do begin
3. C_k=apriori-gen(L_{k-1}); // new candidates
4. forall records $r \in D$ do begin
5. C_r = subset(C_k, r); // candidates contained in r
6. forall candidates $c \in C_r$ do
7. $c.count++$
8. end
9. $L_k = \{c \in C_k | (c.count/size(D)) > minsup\}$
10. end
11. $Answer = L = \bigcup_k L_k$

Fig. 1. Algorithm APRIORI

// Phase 1
Insert into C_k
Select $p.\text{item}_1$,$p.\text{item}_2$,...,$p.\text{item}_{k-1}$,$q.\text{item}_{k-1}$
From L_{k-1} p, L_{k-1} q
Where ($p.\text{item}_1$=$q.\text{item}_1$) and ... and
($p.\text{item}_{k-2}$=$q.\text{item}_{k-2}$) and ($p.\text{item}_{k-1} < q.\text{item}_{k-1}$)
// Phase 2
forall itemset $c \in C_k$ do
forall $(k-1)$-subsets s of c do
if $s \notin L_{k-1}$ then
Delete c from C_k

Fig. 2. Function apriori-gen

APRIORI is based on the fact that $X \supseteq Y \rightarrow Sup(X) \leq Sup(Y)$. Therefore if $Sup(X) < minsup$ then $\forall Y \supseteq X,\ Sup(Y) < minsup$. So we can discard every itemset that has a non large subset.

3 APRIORI-B algorithm

APRIORI-B performs subgroup discovery by learning in several steps a set og classification rules. Given a dataset D, it:

1. removes the $Class$ attribute from D, obtaining $D_{noclass}$;
2. uses APRIORI (described in Section 2.2) on $D_{noclass}$, to obtain the set of large itemsets L;
3. for each itemset $B \in L$ and for each item $H = \{Class = c\}$ where c is a value of the $Class$ attribute, builds the rule $R = B \rightarrow H$;
4. for each rule, computes the rule score metric;
5. sorts rules (in descending order of the metric) and filters them (using a lower bound on the metric $minmetric$ and a maximum number of rules $maxrules$),
6. evaluates the p-value of each rule, by using the randomization test described in Section 3.1.
7. filters the rules, considering a p-value threshold, and obtains the final rule set RS.

APRIORI-B allows the use of several rule evaluation metrics. For the experiment performed in this paper, we used novelty and confirmation (defined in Section 2.1).

Our algorithm is very close to CBA [7] but while CBA uses APRIORI for identifying classification rules with a minimum support, APRIORI-B uses APRIORI in the first learning phase for finding itemsets with a minimum support that are then used as classification rule bodies. Another difference is the following: our algorithm does not aim to build a classifier. Its goal is to find a set of rules that can highlight relations between attributes and the class.

Moreover, one of the main distinguishing features of APRIORI-B is the use of randomization test. The main advantage of using this approach for rule selection is that we obtain immediately a p-value for each rule. This can be very useful to assess rules significance.

3.1 Randomization test

In order to select only the rules having a significant value for the considered metric, we performed a randomization test [9].

First of all, we generated 1000 shuffled dataset, starting from the original one, by independently shuffling the values inside each column. In this way we obtained datasets with the same probabilities for each attribute values but without relations between them. This step was performed before the dataset preparation described in Section 5.1.

We used APRIORI for obtaining the rules, and sorted them using the value of the metric. Then we re-computed the metric for each of the learned rules using all the 1000 shuffled dataset. In this way, for each rule, we obtained a statistical distribution of its metric (i.e. we computed the mean and standard deviation of the metric). By comparing the value of the metric computed by using the original dataset with this distribution, we can assess the significance value of a rule (we considered the values to have a normal distribution).

4 The case study: Marker Analysis

Most common diseases are complex genetic traits [1], where multiple genetic and environmental variables contribute to the observed traits. Because of the multifactorial nature of complex traits, each individual genetic variant (susceptibility allele[3]) generally has only a modest effect, and the interaction of genetic variants with each other or with environmental factors can potentially be quite important in determining the observed phenotype[4]. Genetic association studies, in which the allele or genotype[5] frequencies at markers are determined in affected individuals and compared with those of controls (case-control study design), may be an effective approach to detecting the effects of common susceptibility variants.

The most abundant source of genetic variation in the human genome is represented by single nucleotide polymorphisms (SNPs). SNPs can identify common, but minute, variations that occur when a single unit in a genome sequence (nucleotide) is altered. These variations can be used to track inheritance in families.

Eleven million SNPs of greater than 1% frequency are estimated to exist in the genome and the International HapMap Project has as a primary goal the identification of appropriate sets of tag SNPs that span the genome. These tag SNPs may be able to capture most of the common genetic variants contributing to complex human disease.

At the moment, studies and algorithms able to identify non-random correlations between alleles at a pair of SNPs, have been discussed as a general approach to determine multiple locus involved in human chronic diseases with a genetic component. Moreover, a quantity of "tagging" algorithms for selecting minimum informative subsets of SNPs has recently appeared in the literature.

4.1 Experimental Dataset

As an example of complex genetic trait, we choose Generalized Aggressive Periodontitis (GAP) as case study. Periodonditis is a dental disorder that results from progression of gingivitis, involving inflammation and infection of the ligaments and bones that support the teeth.

The dataset, provided by the Research Center for the Study of Periodontal Diseases, University of Ferrara, collects data from 46 GAP patients (16 males and 30 females) and 102 periodontally healthy control subjects. All subjects were chosen amongst current and permanent residents of the city of Ferrara area. Systemically healthy GAP patients were selected for study among those undergoing periodontal supportive therapy at the Research Center for the Study of Periodontal Diseases, University of Ferrara, and the diagnoses were confirmed by the

[3] Allele: one of several alternative form of a gene or DNA sequence at a specific chromosomal location (locus). At each locus an individual possesses two alleles, one inherited from the father an one from the mother.

[4] Phenotype: the observable attribute(s) of a cell or an individual, brought about by the interaction of genotype and environment.

[5] Genotype: the specific allelic composition of an organism or cell.

same clinician. The clinical diagnosis at the time of the initial visit was based on recent international classification [13]. The periodontally healthy control subjects were selected if they showed no interproximal attachment loss greater than 2 mm at any of the fully erupted teeth. Controls were matched by age and sex with GAP patients. All GAP patients and controls were Caucasian Italian. The study design was approved by the local ethical and written informed consent was provided by all participants in line with the Helsinki Declaration before inclusion in the study.

The following variants in the IL-1 gene cluster have been tested: IL-1α^{+4845} (recorded as $M1$), IL-1β^{+3953} ($M3$), IL-1β^{-511} ($M2$) and also the minisatellite of IL-1RN intron 2 ($M5$). Furthermore, it has been tested a new marker variant at the IL-1F5 ($M6$) gene as described in Scapoli et al. [14]. Besides polymorphisms at IL-1 cluster, other markers have been tested in different pro-inflammatory cytochine such as IL-6 (variant IL-6^{-174} ($M8$) and IL-6^{-622} ($M7$)) and TNF-A (variant TNF-α^{-308} ($M4$)). Finally also a polymorphism at the TNF-α receptor has been tested (TNFRSF1β^{+196} ($M9$)).

4.2 Related Studies

Several studies have shown a role for the involvement of interleukin-1 (IL) gene cluster polymorphisms in the risk of periodontal diseases. In [15] the authors tested polymorphisms, derived from genes of the IL1 cluster, for association with generalized aggressive periodontitis (GAP) through both allelic association and by constructing a Linkage Disequilibrium map of the 2q13-14 disease candidate region. For the IL-1RN intron 2 ($M5$), a statistically significant difference was found between patients and controls in the genotypic distribution, but no significant difference was found for allelic distribution. Authors also observed some evidence for an association between GAP and the IL-1β^{+3953} ($M3$) polymorphism.

For the other IL-1 Cluster polymorphisms, no significant differences were found between patients and controls for both genotypic and allelic frequencies.

Moreover, in [16], the authors showed that allele 1 of the IL-1β^{+3953} ($M3$) and allele 1 of the IL-1RN intron 2 ($M5$) in combination were significantly elevated in GAP as compared to controls.

5 Experiments

5.1 Results on GAP dataset

Dataset preparation The application of the algorithms for subgroup discovery on genetics dataset was performed by an examiner who was blinded as to the correspondence of the M1, M2,. . . ,M9 variables and the related polymorphisms, so that the examiner had not information on previous statistical analyses and on the expected results about IL-1β^{+3953} (M3), IL-1RN (M5) and TNFRSF1β^{+196} (M9) markers and the disease status.

Starting from the blinded dataset originated from the GAP study, we obtained a new dataset on which we ran the experiments. In the original dataset, each marker can assume three possible values: 11, 12 and 22. 11 and 22 are homozygote subjects while 12 define the heterozygote status. As an example, if there are two markers $(M1, M2)$ a possible record of the dataset is $(11, 12)$. In our analysis we consider the configuration of a single chromosome and we want to test, for each marker, whether the allele on that chromosome is 1 or 2. For heterozygote individuals, we do not know on which chromosomes lies the 1: in other words, the allelic configuration for the marker on the two chromosomes could be 12 or 21 with equal probability. The new dataset will contain, for each record from the original dataset all possible configurations of a single chromosome (haplotype) compatible with the record. Therefore, for each record in the original dataset, we generate 2^k tuples in the new dataset, where k is the number of marker analyzed. For example, in the case of the record above, the new dataset will contain the four tuples: $(1, 1)$, $(1, 2)$, $(1, 1)$ and $(1, 2)$.

Results: The dataset obtained (as described in the previous section) was analyzed by using APRIORI-B with two different rule metric, Novelty and Confirmation. The algorithm was configured with the following parameters: *minsup* set to 0.3, *minmetric* set to 0, *maxrule* set to 100 and p-value threshold set to 0.01 .

Rule learned by APRIORI-B using Novelty are shown in Table 2. For each learned rule, the table shows:

- Rule Body, the body of a learned rule containing a conjunction of $Marker = Allele$ tests ;
- State, the disease state associated to the conjunction of $Marker = Allele$ tests in the Rule Body;
- Novelty, the novelty metric value for the rule;
- Rand. Mean, the mean of the novelty values found in the 1000 randomized datasets for the classification rule under analysis;
- Rand. Std, the standard deviation of the Novelty values found in the 1000 randomized datasets for the classification rule under analysis;
- p-value, the rule p-value.

Rules learned by APRIORI-B using Confirmation have not been reported as they are the same learned in the experiment conducted with Novelty even if in a slightly different order.

Analyzing these results, we noticed that some of the rules are related to the two markers that have been reported in literature as involved in the pathology: M3 and M5. The expert confirmed that the correlation between the combination of M3 and M5 found in rule 3 is confirmed by literature [16]. The role of M9 and the combination between M8 and M9, and between M1 and M9 needs further biological investigations.

Table 2. Rule learned by APRIORI-B using Novelty

#	Rule Body	State	Novelty	Rand. Mean	Rand. Std	p-value
1	M9=1	GAP	0.0498	-0.0006	0.0154	0.000530
2	M5=1 M9=1	GAP	0.0408	-0.0006	0.0153	0.003379
3	M3=1 M5=1 M9=1	GAP	0.0383	0.0000	0.0127	0.001288
4	M3=1 M9=1	GAP	0.0377	-0.0001	0.0131	0.001953
5	M8=1 M9=1	GAP	0.0323	-0.0005	0.0127	0.005038
6	M1=1 M9=1	GAP	0.0310	-0.0001	0.0130	0.008699
7	M1=1 M5=1 M9=1	GAP	0.0305	-0.0001	0.0127	0.008153

6 Conclusion And Future Work

In this paper we described a novel algorithms for subgroup discovery named APRIORI-B. This algorithm is based on APRIORI for large itemset generation and randomization test for rule selection.

We developed this algorithm in order to study data obtained from marker analysis. APRIORI-B performance has been evaluated on a real dataset about generalized aggressive periodontitis, and the learned rules were judged interesting by the biologist.

Given this set of rules, further investigation could be made identifying the group of patients which present the marker combination specified by one of the rules. The comparison of the clinical state of these patient groups can be useful to conduct a more specific study of the disease (e.g. finding different disease phenotypes). This will be matter of future works. Moreover, a new dataset about sclerosis will be analyzed.

7 Acknowledgments

This work has been partially supported by NOEMALIFE under the "SPRING" regional PRRITT project, by the PRIN 2005 project "Specification and verification of agent interaction protocols" and by the FIRB project "TOCAI.IT".

References

1. Lewin, B.: Genes VII. Oxford University Press, Oxford (2000)
2. Klösgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and Data Mining. AAAI/MIT (1996) 249–271
3. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In Komorowski, H.J., Zytkow, J.M., eds.: PKDD. Volume 1263 of Lecture Notes in Computer Science., Springer (1997) 78–87
4. Wrobel, S.: Inductive logic programming for knowledge discovery in databases. In Dzeroski, S., Lavrac, N., eds.: Relational Data Mining. Springer (2001) 74–101

5. Kavsek, B., Lavrac, N., Jovanoski, V.: Apriori-sd: Adapting association rule learning to subgroup discovery. In Berthold, M.R., Lenz, H., Bradley, E., Kruse, R., Borgelt, C., eds.: IDA. Volume 2810 of Lecture Notes in Computer Science., Springer (2003) 230–241
6. Lavrač, N., Kavček, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research **5** (2004) 153–188
7. Liu, B., Hsu, W., Ma, Y.M.: Integrating classification and association rule mining. In: KDD-98. (1998) 80–86
8. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In Bocca, J., Jarke, M., Zaniolo, C., eds.: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Morgan Kaufmann (1994) 487–499
9. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall, London (1993)
10. Lavrac, N., Flach, P., Zupan, B.: Rule evaluation measures: A unifying view. In Dzeroski, S., Flach, P.A., eds.: ILP. Volume 1634 of Lecture Notes in Computer Science., Springer (1999) 174–185
11. Flach, P.A., Lachiche, N.: Confirmation-guided discovery of first-order rules with tertius. Machine Learning **42**(1/2) (2001) 61–95
12. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Comput. Surv. **38**(3) (2006) 9
13. Tonetti, M.S., Mombelli, A.: Early-onset periodontitis. Ann Periodontol **4** (1999) 39–53
14. Scapoli, C., Tatakis, D., Mamolini, E., Trombelli, L.: Modulation of clinical expression of plaque-induced gingivitis: interleukin-1 gene cluster polymorphisms. J Periodontol **76** (2005) 49–56
15. Scapoli, C., Trombelli, L., Mamolini, E., Collins, A.: Linkage disequilibrium analysis of case-control data: an application to generalized aggressive periodontitis. Genes Immun **6** (2005) 44–52
16. Parkhill, J.M., H.B.C.I.H.P., Taylor, J.: Association of interleukin-1 gene polymorphisms with early-onset periodontitis. J Clin Periodontol **27** (2000) 682–689