A Framework for the Application of Decision Trees to the Analysis of SNPs Data

Linda Fiaschi, Jonathan M Garibaldi and Natalio Krasnogor

Abstract— **Data mining is the analysis of experimental datasets to extract trends and relationships that can be meaningful for the user. In genetic studies these techniques have revealed interesting findings, especially in the heritable predisposition to contract specific diseases. One of these diseases which is still under extensive analysis is pre-eclampsia, a progressive disorder which occurs during pregnancy and soon after the birth, affecting both the mothers and their babies. There are many choices to be made in the application of the various data mining techniques that may be used to study general genotype-phenotype associations. The aim of this paper is to describe the general framework that we adopted in the application of decision tree algorithms to the analysis of SNPs data related to cases of pre-eclampsia. The results show the validity of this methodology to detect a subset of attributes associated with the predictable variable, providing a reduction in the size of the dataset. Moreover, from the clinical point of view, it confirmed the medical interpretation of the 'corrected birth-weight centile' (CBC) value of 10 being a meaningful cut-off and confirmed association between an infant's CBC and the 'week of delivery' parameter. We hope that the generic framework described here will be of use to other researchers analysing such data.**

I. Introduction

Data mining is an analytic process designed to explore (often large amounts of) data in search of consistent patterns or systematic relationships between variables; the findings may then be validated by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction. The initial exploration of the data involves data cleaning, data integration, data transformation and record subsets selection. Following this, advanced computational and statistical methods are applied in order to extract the most interesting, novel, useful and valid data patterns from the given datasets. Finally, visualization and knowledge representation techniques are used to present the extracted knowledge to the user [1], [2], [3].

Data mining has been commonly applied to the field of genetic analysis [4]. The study of the human genome has become one of the most challenging goals for scientists. All the biological information needed to build and maintain a living example of an organism are contained in a double-stranded molecule consisting of two chains running in opposite directions and called deoxyribonucleic acid (DNA) [5]. DNA is essentially a sequence of four types of molecules called 'bases', labeled 'C', 'G', 'A' and 'T', which join into complementary 'base-pairs' ('C' with 'G', and 'A' with 'T'). Human DNA is estimated to comprise around 3 billion base-pairs, of which around 99.9% are the same — there is only a small percentage that makes the difference between individuals [6]. While at most positions in human DNA the same base is found, approximately once every 100 to 300 bases a different base may be found. Such an alteration is called a Single Nucleotide Polymorphism (SNP). The majority of these changes have no effect or at least not yet known; but others can cause subtle differences in physical or psychological characteristics. Some of them may actually affect a person's response to drug therapy and even confer a personal susceptibility or resistance to a certain disease, determining then the severity or progression of it. For this reason, analysis of SNPs has become the subject of extensive research [7], [8], [9]. Within the diseases considered to be related to genetic causes, there is one called pre-eclampsia (PE) which is currently under genetic analysis for any heritable association [10], [11], [12], [12], [13], [14], [15]. PE is a progressive disorder which occurs during pregnancy and in the period soon after the birth and it affects both the mother and the baby. The major symptoms are high blood pressure, swelling, proteins in the urine and problems with vision. It occurs in around 5-8% of all pregnancies and, together with other disorders of high blood pressure during pregnancy, it is responsible globally for an estimated 76,000 maternal and 500,000 infant deaths each year [16].

There are different models that have been used by researchers for studying general genotype-phenotype associations depending on the kind of application. Population-based, family-based strategies and their numerous extensions are all widely used to detect genes associated with complex diseases. This paper is exclusively focused on association studies, defined as "a gene-discovery strategy that compares allele frequencies in cases and controls to assess the contribution of genetic variants to phenotypes in specific populations" [17]. This kind of study implies the creation of two different groups among the population. One of the groups is composed of 'cases' (people with a disease or a condition) and the other one is composed of 'controls' (individuals without this condition). Extracting features from these two different groups and comparing them with each other gives the possibility to detect classification rules in a straight forward way. In PE, for instance, a population of mothers can be considered and they can be split into sick mothers and healthy ones. However, the problem can also be studied by considering different prediction variables, like for instance a clinical feature of the disease.

Within the general data mining tools, there is a sub-class of algorithm widely used for case-control analysis in SNPs studies: the decision tree algorithms [18], [19], [20]. These

The authors are with the School of Computer Science, University of Nottingham, Wollaton Road, NG8 1BB, UK.

are based on classification trees to predict membership of cases in the classes of a categorical dependent variable. In the study shown in this paper, three of these algorithms are taken in consideration: ID3, ADTree and C4.5 [21], [22], [23], [24], [25]. The aim of this paper is to compare or contrast the results obtained from a variety of decision tree algorithms in order to identify commonality between trees. A full comparison of a wide variety of approaches, while interesting, is outside the scope of this paper.

ADTree is a natural generalization of decision trees in which rules are usually smaller in size and easier to interpret compared to the other boosted decision tree algorithms. ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples using the concept of information entropy. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the n partitioned subsets (where n is the number of possible values of an attribute) to get their 'best' attribute. The attributes must have a fixed number of values and the class must be discrete as well. C4.5 improves on ID3 as it can handle with both continuous and discrete attributes and training data with missing attribute values. Moreover C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

Before applying these algorithms, a complex pre-processing stage of the initial database is performed in order to encode attributes (where necessary), explore the data, treat missing and unbalanced data, and set parameters. In the following Section of this paper the proposed methodology is shown through the two main streams of action: pre-processing and proper analysis of the dataset. The application of this technique to an example of a medical database containing heterogeneous information about a list of patients affected by pre-eclampsia is described and the results are shown. The features of each individual comprise both genetic and clinical data. The final goal of this research is both to propose a valid method for SNPs analysis and, from the medical point of view, to discover any possible association, either genetic or phenotypic, with the specific disease.

II. Methodology

This is a kind of progressive analysis through which significant results are detected in the first stage then deepened and possibly confirmed in the subsequent steps. All the stages are explained in this paragraph step by step in order to allow the methodology to be fully described, as shown in Figure 1.

A. Database Pre-processing

In this paper we will show the analysis related only to a specific dataset. From now on, we refer to a database (DB) as a specific subset of records obtained from the original (entire) set of records. In general, we consider the original set of data to consist of one or more attributes of SNPs and one or more attributes of phenotypic information.

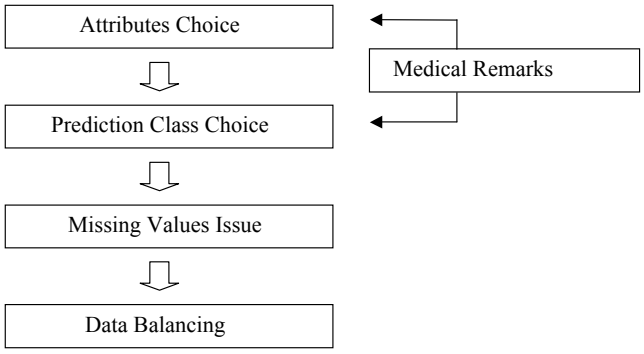

Fig. 1. Sequence of steps to follow in the pre-processing of a general dataset composed of medical and SNPs attributes.

1) Choice of Attributes: Every time a new DB is created there are attributes that may be deleted as considered useless or not informative for that specific population. In the first stage, all the remaining attributes can be kept in order to detect any possible features that are significant in this analysis. The results which are further obtained will remove the less informative attributes. Considering the final DB, different kind of analysis can be performed in this study: some of them are over the whole attributes, other are only over the SNPs and others over phenotypic attributes.

Sometimes within SNPs analysis, the set of data may include information about families. In this case, there can be some features coming from the relatives of the individuals under analysis (mothers or babies in the example of this paper). This information can be easily transferred from the columns of one row to additional columns of a given individual, and hence analyzed as a new attribute. For instance, if we consider a dataset of only babies, we could add new columns with the genetic information about their parents.

2) Choice of Predictive Class: In general it is interesting to analyze the data considering different predictive variables for the same population. Many decision tree algorithms require the prediction class to be a categorical attribute and, more specifically, a Boolean one. If this is not the case, a threshold needs to be found in order to transform the variable into a Boolean one. As this analysis is a case-control one, we consider only Boolean prediction variables.

3) Consideration of Missing values: Some decision tree algorithms can deal directly with attributes containing missing values, but others cannot. It may be necessary to eliminate the missing values or to adopt another strategy such as imputation of missing data. It may not be appropriate to simply remove rows with missing values, in case there are many missing values and their deletion could therefore affect significantly the size of the DB. Attributes containing many missing values may still be dropped at a later stage. On the other hand, in the specific application in this paper it is possible to keep the missing values as the algorithms chosen can deal with a codification for them.

4) Balancing of Data: The balancing of the DB is quite an important issue to be considered before performing the analysis. Often, the ideal situation is to have approximately half of the individuals belonging to the cases and half to the

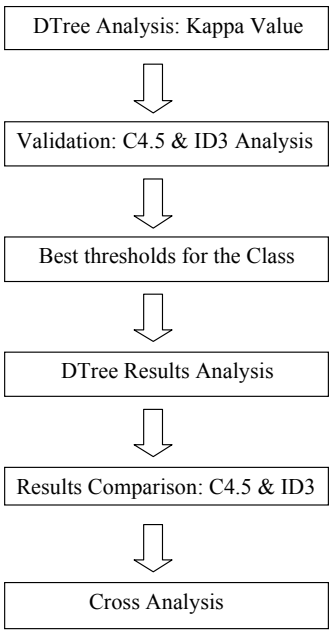

Fig. 2. Sequence of steps to follow in the analysis of a genetic and clinical DB with the CBC as predictable variable. The applied algorithms are: ADTree, Id3 and C4.5.

controls in order to have the least biased performance. If this is not the case, it is possible to create a new DB by selecting randomly a fixed number of people from both the groups.

5) Medical remarks: As we are analyzing clinical data it is useful to have feedback from the medical side throughout the whole process of analysis. There are attributes and prediction classes that have more relevance than others for the user, the doctor in this case. The final thresholds found for the prediction variable need to be considered from the medical point of view before defining them to be interesting. In general, any finding that can be significant from the statistical point of view it is not always meaningful for the medical community.

B. Database Analysis

There are different kinds of algorithms for data mining that can be used but the essential idea is to perform repeated case-control analysis, each time defining the class and determining the subset of the original database to use. In this paper, three decision tree algorithms which all work with a nominal class were used: ADTree [26], ID3 and C4.5.

The steps for the data analysis process are shown in Figure 2. For this paper, the Weka software [21] was used to perform the analysis of the DB (note that in Weka the C4.5 algorithm is known as J48).

1) ADTree Analysis - Kappa value: In the first step the DB is analyzed with one of the three mentioned algorithms, arbitrarily chosen. In this paper, for instance we will start with the ADTree algorithm. If the chosen predictive variable is a continuous one, it needs to be converted to a Boolean attribute. Therefore, a range of thresholds has to be chosen for this class in order to detect the ones which give the most significant results. It is possible to select a fixed number of different thresholds, for instance 10, within the range of the variable and check for the reliability of the results.

All algorithm parameters are set to their default value (in Weka), as an exhaustive examination of the effects of varying parameters is outside the scope of this paper.

The validity of the analysis has been calculated through the use of the Kappa statistic, defined as the proportion of agreement corrected for chance between two judges assigning cases to a set of categories [27]. Although boundaries are arbitrary, we consider $K > 0.20$ a fair agreement in accordance with [28].

From the results it is easy to establish which thresholds provide a statistically significant result according to a Kappa value greater than 0.2. Then a further selection can be done by means of clinical feedback; there may be thresholds which don't have any particular medical meaning and other ones which correspond to medically accepted values.

In this analysis ten-fold cross-validation is used and repeated ten times with different seeds to create different random partitions of the data. The seeds can be chosen either arbitrarily or randomly. An average of the obtained results is then calculated.

2) Validation - C4.5 and ID3 analysis: In order to confirm (or not) these findings, the DB is processed with two other decision tree algorithms, C4.5 and ID3, with the same thresholds used in the previous analysis. These algorithms are also run with a ten-fold cross-validation using different seeds and the average of the Kappa value obtained from each result is calculated.

3) Determining Threshold for the Predictive Variable: In the next stage the focus is only on the subsets whose thresholds give significant results (the ones for which Kappa is greater than 0.2) in order to check the Kappa trend. The algorithms that give us the best results are run again and this time with the subsets of data whose class thresholds is included in the range previously found, choosing a number of different threshold values, for instance 10 or 20. From the results it is possible to detect the subset whose thresholds give a Kappa greater than 0.2.

Once that the thresholds have been chosen it is important to check the number of individuals involved in each test and the ratio of cases to controls in order to deal with a reliable test. If one of the final subsets has a case-control ratio above or below 50%, it cannot be taken in consideration for further analysis. Lots of studies have been published about the optimal case-control ratio and size of the dataset depending on the kind of application used in the analysis [29], [30], [31], [32]. In this case, it is reasonable to consider a limit for the cases-controls ratio around one-third to be an acceptable one, as far as the amount of cases and controls don't fall below an arbitrary threshold of around 100 individuals.

4) ADTree Results Analysis: Focusing on those DBs which give the best Kappa, it is now possible to analyze the results of the tests in order to determine whether they can be considered reliable. The first step consists of the comparison of the decision trees obtained from the DBs processed with ADTree, each one with a different class threshold. It can

happen that, comparing the different trees, they have different shape and therefore different rules of classification but it is still possible to list the attributes which are present in all the trees obtained. The attributes can be detected from each node of the final classification tree.

5) Results Comparison - C4.5 and ID3: The same procedure is applied also to the second and the third algorithm, as far as they provide reliable results. For instance, for the set of trees obtained with ID3, each one with a different class threshold, we make a list of the attributes which are present in all the obtained trees. In the end we will get three lists of the attributes common to different class thresholds, each one from a different algorithm. Comparing these lists, we can find the SNPs which are detected from different algorithms and therefore which are present in all the results of this analysis.

6) Cross analysis: A cross analysis can now be performed between the decision trees obtained with the best algorithms, considering each time the same thresholds. For instance we can consider the ADTree trees obtained with specific class threshold and compare them with the respective ones obtained from the ID3 algorithm. As before, we can make a list of the attributes present in the results from different algorithms but with the same threshold. If there is a SNP which appears in all the lists created, it is more likely to have a reliable association with the predictable variable.

III. Experimental Results

In this section an example of the application of this methodology to a real dataset related to PE is described.

A. Experimental Data

The DB under analysis contains 4529 instances and 105 attributes. The original dataset is composed of mothers, babies, fathers, grandparents and other relatives of the baby; there are fifty-two genetic attributes (SNPs) split across seven genes and fifty-three phenotypic (clinical) attributes, as follows:

(1) Genotype: 52 attributes:
- AGT gene: SNPs 1-8, alleles 1 and 2
- AGTR1 gene: SNPs 9-12, alleles 1 and 2
- TNF gene: SNPs 13-16, alleles 1 and 2
- F5 gene: SNP 17, alleles 1 and 2
- NOS3 gene: SNPs 18-22 and 24, alleles 1 and 2
- MTHFR gene: SNPs 25, 26, alleles 1 and 2
- AGTR2 gene: SNP 27

(2) Phenotype: 53 clinical attributes
- 5 concerning the individual's identity;
- 34 concerning maternal data, such as physical and physiological parameters, pregnancy details and current treatments;
- 6 concerning fetal data, such as the weight and gestational age at birth;
- 8 concerning the medical history of parents, partners or siblings of affected mothers.

The individuals of most interest for this disease are the mothers and the babies. There are actually four different conditions present in the original database: pre-eclampsia, eclampsia, other hypertensive diseases and normotensive (normal blood pressure). The only condition which is investigated in this paper is pre-eclampsia.

TABLE I

Prediction attributes for the babies

Attributes for the Babies	Type	Range
CBC	Percentage	$1-100$
Delivery gestation week	Integer	$22-42$ weeks

TABLE II

Prediction attributes for the mothers

Attributes for the mothers	Type	Range
CBC	Percentage	$1-100$
Delivery gestation week	Integer	$22-42$ weeks
Sys/Dias Pressure Post Partum	Integer	$87-178$
Highest Systolic	Integer	$101-200$
Highest Diastolic	Integer	$65-150$
Highest Proteinuria	Real	$0.24-32.03$
Highest ALT	Integer	$2-875$
Highest Urate	Integer	$50-812$
Highest Creatinine	Integer	$49-990$
Highest Urea	Real	$1.6-33.8$
Lowest Platelets	Integer	$12-443$

B. Data Base Pre-processing

From the initial Database a subset is created containing only babies born from mothers with pre-eclampsia.

1) Attributes: In the first stage most of the attributes are kept. There are only a few attributes which are not meaningful when we consider a database composed of babies. These are the mothers features, such as blood pressure and blood test results.

2) Predictive Class: The idea is to analyse the data considering different prediction variables as shown in Table I and in Table II.

One of the most interesting variables listed in these tables is the 'corrected birth-weight centile' (CBC). This is the value of the weight of the baby at birth (as a percentage of the population) corrected for gestational age at birth, baby sex, ethnicity, mother's height, mother's weight and number of pregnancies. Hence, a CBC of 50 is the normal weight at birth, below this threshold it is considered underweight and a CBC exceeding this threshold is considered overweight. For each of these outputs we can decide different thresholds to define the cases and the controls in the dataset in order to perform a case control analysis. For instance we could choose the following values: CBC $= 50$, Delivery gestation $= 35$, Systolic Pressure post partum ≤ 140, or Diastolic Pressure post partum ≥ 90

The results shown in this paper are from a DB consisting only of babies, created from the original one by deleting the attributes considered not informative for a population of babies. The CBC attribute has been chosen as the predictive class and the final DB consists of 372 babies and 58 attributes. Beside the 53 SNPs listed above, there are six clinical variables for the babies: 'Fetal disease status', 'Gestation at birth (weeks)', 'Gestation at birth (days)', 'Weight of the infant', 'Live at birth' and CBC.

3) Missing values: Different trials were performed in order to understand if it is informative to retain the missing values or if their removal could have improved the study. We applied the algorithms to a dataset cleaned from the missing values and to a dataset with the missing values retained. The results obtained were the essentially unaltered, indicating that (for this data set) we can retain the missing values using the appropriate codification for the chosen algorithm.

4) Balancing of the data: As the CBC class is not Boolean, at this point it is not possible to balance the data because it is not yet clear the amount of cases and controls. Balancing of the data can be performed later, when a fixed CBC threshold is chosen and therefore the babies with a CBC greater than that threshold are considered as controls and these with CBC below that threshold considered as cases.

C. Data Base Analysis: Babies with pre-eclampsia

1) ADTree Analysis: In the first stage the DB is analyzed with the ADTree software from Weka. A range of thresholds has been chosen for the CBC class in order to detect the ones which give the significant results. There are 9 different thresholds, from a CBC of 10 to a CBC of 90, and for each the Kappa value is calculated as shown in Table III.

From Table III it is clear that the first three thresholds (CBC of 10, 20, and 30) provide a statistically significant result (Kappa $>$ 0.2) whereas the others have a quite low Kappa value. The ADTree algorithm is then run again with a set of 9 different seeds and the average of the result has been calculated as shown in Table IV.

2) Validation — C4.5 and ID3 Analysis: In order to have a validation of these findings, the DB has been processed with the other two decision tree algorithms. The thresholds are the same used in the previous analysis. These algorithms have been run nine times with different seeds and the average of the Kappa value has been calculated as shown in Table IV.

From these results of Table IV it is clear that with C4.5 a result similar to ADTree has been obtained with a significant Kappa for thresholds of 10, 20 and 30. Regarding ID3, no significant results have been obtained over the thresholds as shown in Table IV. Thus, ID3 does not appear to be able to detect relevant findings in this application, we did not investigate this limitation any further.

3) Best Threshold(s) for the Predictive Variable: As the CBC thresholds of 10, 20 and 30 have shown to be relevant, in the next stage the interest is focused only on the CBC range between 4 and 30, in order to detect any other threshold with a good Kappa. The two algorithms that give us the best results are run again and this time with 14 different thresholds of CBC in the range $4 - 30$.

From examination of Figure 3 it can be seen that the three thresholds with the best Kappa value are 6, 10 and 28. The fact that the trend in Kappa is not monotonic with CBC may be due to the presence of noise in the data but could also be due to the complex correlation between attributes such as CBC and week of delivery (later discovered). Once we have fixed the CBC values, we can keep into consideration balancing of the data. The number of cases and controls involved in each test results are shown in Figure 4. In particular:

- for CBC = 6: 147 cases (39.5%) and 225 controls
- for CBC = 10: 177 cases (47.6%) and 195 controls
- for CBC = 28: 243 cases (65.3%) and 129 controls

These results are acceptable regarding both the absolute size of the population (372) and the proportions of cases and controls, as the case-control ratio is above 0.33 and there are more than 100 individuals for each group.

TABLE V
COMMON ATTRIBUTES TO THE THREE CBC THRESHOLDS 6,10 AND 28 FOR THE ADTREE ALGORITHM.

			CBC			
Gene	SNP	Allele	6	10	28	All
AGT	1	1	y			
AGT	3	2			y	
AGT	6	2			y	
AGTR1	10	2		y		
AGTR1	11	2			y	
AGTR1	12	2	y			
F5	17	2	y	y		
NOS3	19	2		y		
NOS3	21	2	y	y	y	y
NOS3	24	2	y	y		
MTHFR	26	2			y	
AGTR2	27	2	y	y	y	y

4) ADTree Results Analysis: Focusing on these three DBs, the next step consists of the comparison of the three decision trees obtained from the three DBs processed with ADTree. Comparing the three different trees it is clear that they have different shape and therefore different rules of classification but it is possible to list the attributes common to all of them. Besides 'gestational week at birth'(which is always present), the attributes found for CBC equal to 6,10 and 28 are shown in Table V. In the last column, the attributes common to the first three columns are shown.

5) Results Comparison — C4.5 and ID3: The same procedure is applied also to the C4.5 algorithm as it provided similar results. The list of the SNPs are shown in Table VI. Concerning the clinical variables, there is still the attribute 'Gestational week at birth' which is common to the algorithm results and the variable 'sex' which is present only in the first two thresholds (CBC = 6 and CBC = 10) .

6) Cross Analysis: A cross analysis can now be performed between two decision trees obtained with the two algorithms (ADTree and C4.5), considering each time the same thresholds (CBC = 6, 10, 28). The results are shown in Table VII.

Furthermore, if we focus our attention on the results when the CBC is 28 we can create a new dataset composed of the common attributes found in the results from both the two algorithms. These attributes are: 'sex' , 'Gestational week at birth', AGT SNP3, AGTR1 SNP11 and NOS3 SNP 21. Processing this new dataset with both the ADTree and C4.5 algorithm, we find two interesting rules which are common to the two final trees, with a statistical significance of $k = 0.38$ for C4.5 and $k = 0.41$ for ADTree. The first rule claims that male babies, born after the 35$^{\text{th}}$ week of gestation and with an AGT SNP3 allele2 of 1 have a good probability

TABLE III

STATISTICAL RESULTS FROM ADTREE: CBC=10-90

CBC Thresholds	10	20	30	40	50	60	70	80	90
Kappa	0.35	0.23	0.20	0.02	−0.03	0.02	−0.01	0	−0.01

TABLE IV

KAPPA AVERAGED OVER NINE RUNS FOR ADTREE, C4.5 AND ID3 ALGORITHMS

CBC Thresholds	10	20	30	40	50	60	70	80	90
ADTree Kappa	0.38	0.32	0.29	0.18	< 0.07	< 0.04	< 0.04	< 0.05	< 0.04
C4.5 Kappa	0.27	0.22	0.28	0.18	< 0.17	< 0.18	< 0.05	0	0
ID3 Kappa	0.15	0.14	0.18	< 0.09	< 0.16	< 0.11	< 0.12	< 0.11	< 0.05

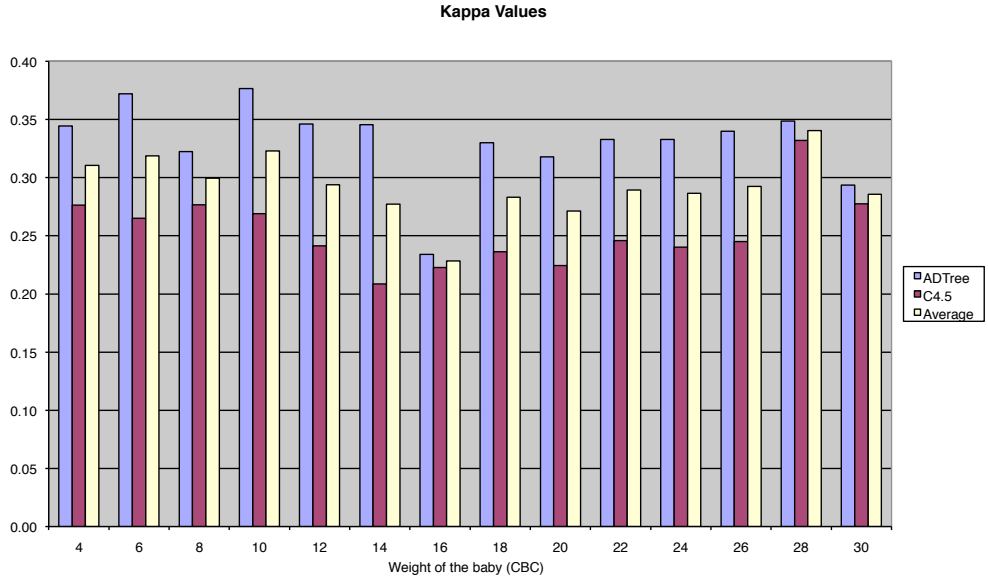

Fig. 3. Kappa Values of the two applied algorithms (ADTree and C4.5) versus weight of the baby expressed as CBC within the range CBC= 4-30.

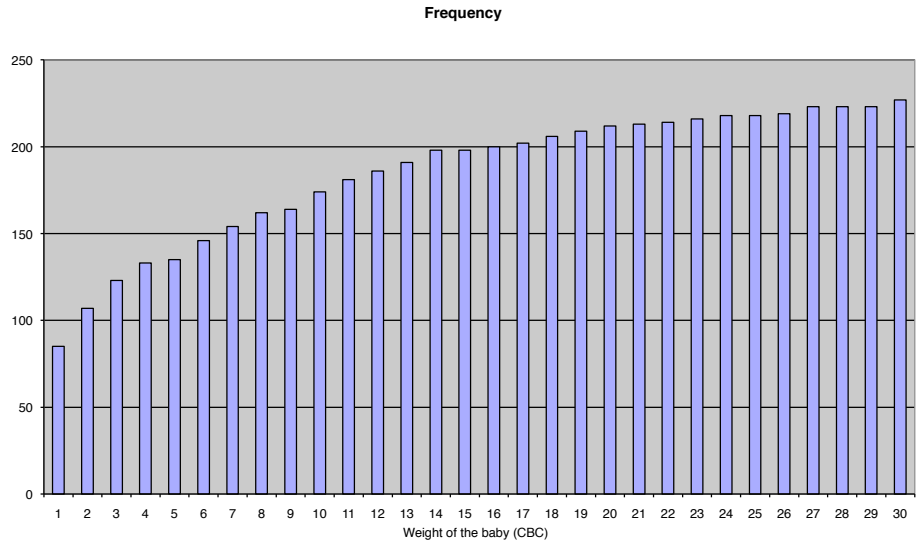

Fig. 4. Number of cases versus the CBC of the babies.

TABLE VI

COMMON ATTRIBUTES TO THE THREE CBC THRESHOLDS 6,10 AND 28 FOR THE C4.5 ALGORITHM.

			CBC			
Gene	SNP	Allele	6	10	28	All
AGT	1	1		y	y	
AGT	1	2		y	y	
AGT	3	2	y	y	y	y
AGT	4	2	y			
AGT	6	1			y	
AGT	7	2			y	
AGT	8	1	y			
AGT	8	2	y	y	y	y
AGTR1	9	1	y	y	y	y
AGTR1	9	2	y	y		
AGTR1	10	1		y	y	
AGTR1	10	2	y			
AGTR1	11	1	y			
AGTR1	11	2	y		y	
AGTR1	12	1		y	y	
AGTR1	12	2	y	y	y	y
TNF	13	1	y	y		
TNF	13	2	y	y		
TNF	14	2	y	y	y	y
TNF	15	2	y			
TNF	16	1	y	y		
TNF	16	2			y	
F5	17	2		y	y	
NOS3	18	2		y		
NOS3	19	1	y			
NOS3	19	2	y	y		
NOS3	20	1	y	y	y	y
NOS3	20	2	y			
NOS3	21	1			y	
NOS3	21	2	y		y	
NOS3	22	1			y	
NOS3	22	2	y	y	y	y
NOS3	24	1	y	y	y	y
NOS3	24	2	y			
MTHFR	25	1	y	y	y	y
MTHFR	25	2	y	y	y	y
MTHFR	26	2	y	y		
AGTR2	27	1		y	y	
AGTR2	27	2	y	y		

to have a normal weight ($CBC > 28$). The confidence of the C4.5 algorithm is measured by the ratio between the corrected classified instances over the uncorrected ones, which is 84/24. The ADTree measure of confidence is instead made by the 'classification margin', analyzed on prior work [33] and it has a absolute value of 1.29. The second finding shows that male babies, born after the 35th week of gestation and with an AGT SNP3 allele2 of 2 and an AGTR1 SNP11 allele2 of 1 have a good probability to be under weight ($CBC < 28$). For the C4.5 the confidence parameters measures 21/5 and for the ADTree the classification margin has an absolute value of 0.76.

Following these results, we have performed the analysis with only one attribute, the 'delivery gestation week', and the CBC predictive variable. We find out that there is an association between these two parameters with a good significance as shown by the Kappa value of 0.4212 for both the ADTree and C4.5 analysis. The ADTree algorithm detects an interesting threshold for the 'GestationatBirthw' equal to 35.5 to discriminate the small babies (cases) from the normal one (CBC > 10) and in C4.5 the threshold is set at 35 weeks of pregnancy. This means that babies delivered before 35 or 35.5 week of gestation are likely to have a CBC < 10.

TABLE VII

COMMON ATTRIBUTES TO THE THREE CBC THRESHOLDS 6,10 AND 28 FOR THE TWO ALGORITHMS: ADTREE AND C4.5.

Gene	SNP	Allele	CBC 6	CBC 10	CBC 28
AGT	3	2			y
AGTR1	11	2			y
AGTR1	12	2	y		
F5	17	2		y	
NOS3	19	2		y	
NOS3	21	2			y
AGTR2	27	2	y	y	

IV. CONCLUSION

The methodology shown in this paper provides researchers with a guideline for data mining in the specific application of case-control analysis for SNPs. This technique may find an association between the SNPs and the disease or its phenotypes. However, it is also possible that the results don't show a significant direct connection between the SNPs and the disease as found in this study. In this case it is still possible to detect a reduced number of SNPs that may play an important role in the genetic association, as for example in this specific experiment.

From the methodological point of view, we conclude that thanks to this strategy, some attributes are rejected as not relevant for the analysis, the number of the instances are decreased and a set of attributes, clinical or genetic, are found to be correlated to the predictive variable, as show in the lists of the common attributes in the example described in this paper, see Table V, Table VI and Table VII. From a comparison of these Tables it is also clear that there are SNPs such as AGT 2 and AGT 5 that never appear in the results; these SNPs can thus be ignored in further analysis.

From the clinical perspective, there are at least two important findings which emerge from this methodology. The first is the significance of the threshold CBC of 10. From the study on the validity of the thresholds three different values have been found: i.e. 30, 20 and 10. The feedback from the medical point of view confirmed the clinical importance of a CBC of 10 for babies affected by pre-eclampsia, as it is a clinically accepted threshold used to identify growth restricted babies, which have then a higher risk of problems in the neonatal period. The second finding is the dependency of the CBC on the 'week of delivery' parameter. In the formula for calculating the CBC, the birth weight is adjusted considering parameters including the 'week of delivery'. This means that there shouldn't be any association between these two attributes. From the results of this analysis on PE disease, an association between these two parameters has been found: women with pre-eclampsia who deliver before 35 weeks of pregnancy are more likely to give birth to babies with a CBC under the value of 10.

The proposed methodology provides (besides the opportunity to find new and challenging results) a useful tool for the

screening stage where a reduction in the number of cases is the main goal.

V. Future Work

An important observation arises concerning the significance of considering the genotype of the mothers rather that the babies. It is still difficult and risky to collect information related to the DNA of babies in pregnancy, as this requires an invasive test. On the other hand we can easily collect such information from the mothers. For this reason, in further research, the analysis may be focused on only the mothers, using the CBC of the baby as the predictive variable.

A second consideration concerns the re-codification of the SNPs in the DB. The whole human DNA chain is divided in 23 different pairs of chromosomes, each one therefore composed by two copies called alleles. The SNP is encoded by two numbers, one for each allele (for instance 1/2). A set of alleles at different places that are present in the same chromosome is called a haplotype [17]. When we perform an analysis between more than one SNP, we consider two haplotypes each one coming from each chromosome pair. It is then important to know which allele comes from a specific chromosome pair. That means it is relevant to know whether the SNP is 1/2 or 2/1. If this information is not available, as in most cases, we have to encode the SNPs considering a SNP 1/2 to be the same as a SNP 2/1.

A limitation of the current study is the redundant interaction between attributes, ignored by these algorithms. A possible selection and elimination of the attributes which are very related to each other can be performed in a preprocessing stage as future work.

The last remark concerns the clinical condition of the grandmothers when their mothers were born. A heritable trend can be detected across the two generations validating the genetic association of this disease. Unfortunately, the current database doesn't contain sufficient information to perform this kind of analysis.

Acknowledgments

We would like to thank Dr. Linda Morgan and Dr. Kevin Morgan from Clinical Chemistry Division, Institute of Genetics at the Queen's Medical Center, Nottingham.

This study was supported by the BIOPTRAIN FP6 Marie-Curie EST Fellowship (MEST-CT-2004-007597).

References

[1] D. Hand, H. Mannila, and P. Smyth, "Principles of data mining," *MIT Press*, pp. 385–394, 2001.

[2] W. Frawley, G. Piatetsky-Shapiro, and C. Matheus, "Knowledge discovery in databases: An overview," *AI Magazine*, pp. 213–228, 1992.

[3] J. Han and M. Kamber, *Data Mining: Concept and Techniques*. Morgan Kaufmann, 2006.

[4] B. S. Weir, *Genetic data analysis II*, 2nd ed. Sunderland, Massachuttes: Sinauer Associated, Inc., 1996.

[5] [Online]. Available: http://www.ncbi.nlm.nih.gov/About/primer

[6] [Online]. Available: http://www.genetics.gsk.com/overview.htm

[7] I. C. Gary, D. A. Campbell, and N. K. Spurr, "Advances in knowledge discovery and data mining," *Human Molecolar Genetics*, vol. 9, no. 16, pp. 2403–2408, 2000.

[8] N. Schork, D. Fallin, and J. Lancbury, "Single nucleotide polymorphism and the future of genetic epidemiology," *Clinical genetics*, vol. 58, pp. 250–264, 2000.

[9] K. Lohmueller, C. Pearce, M. Pike, E. Lander, and J. Hirschhorn, "Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease," *Nat. Genet.*, vol. 33, pp. 177–182, 2003.

[10] S. Daher, N. Sass, L. Oliveira, and R. Mattar, "Cytokine genotyping in preeclampsia," *Am J Reprod Immunol.*, vol. 55, pp. 130–135, 2006.

[11] A. Fekete, A. Ver, K. Boegi, A. Treszl, and J. Rigo, "Is preeclampsia associated with higher frequency of HSP70 gene polymorphisms?" *Eur J Obstet Gynecol Reprod Biol*, vol. 126, pp. 197–200, 2006.

[12] E. Kamali-Sarvestani, S. Kiany, B. Gharesi-Fard, and M. Robati, "Association study of il-10 and ifn-gamma gene polymorphisms in iranian women with preeclampsia," *J Reprod Immunol.*, vol. 72, pp. 118–126, 2006.

[13] G. Kobashi, K. Shido, A. Hata, H. Yamada, E. Kato, M. Kanamori, S. Fujimoto, and K. Kondo, "Multivariate analysis of genetic and acquired factors; t235 variant of the angiotensinogen gene is a potent independent risk factor for preeclampsia," *Semin Thromb Hemost.*, vol. 27, pp. 143–147, 2001.

[14] J. Lin and P. August, "Genetic thrombophilias and preeclampsia: a meta-analysis," *Obstet Gynecol.*, vol. 105, pp. 182–192, 2005.

[15] R. Skjaerven, L. J. Vatten, A. J. Wilcox, T. Ronning, L. M. Irgens, and R. T. Lie, "Recurrence of pre-eclampsia across generations: exploring fetal and maternal genetic components in a population based cohort," *Obstet. gynecol. surv.*, vol. 61, pp. 162–163, 2006.

[16] [Online]. Available: http://www.preeclampsia.org

[17] N. M. Laird and C. Lange, "Family-based designs in the age of large-scale gene-association studies," *Nature Reviews Genetics*, pp. 385–394, 2006.

[18] D. H. Kim, S. Uhmn, Y. W. Ko, S. W. Cho, J. Y. Cheong, and J. Kim, *Computational Science and Its Applications ICCSA 2007*. Springer Berlin / Heidelberg, 2007, vol. 4707, ch. Chronic Hepatitis and Cirrhosis Classification Using SNP Data, Decision Tree and Decision Rule, pp. 585–596.

[19] K. Y. Liu, J. Lin, X. Zhou, and S. T. Wong, "Boosting alternating decision trees modeling of disease trait information," *BMC Genet.*, vol. 6, 2005.

[20] Q. Xie, L. D. Ratnasinghe, H. Hong, R. Perkins, Z.-Z. Tang, N. Hu, P. R. Taylor, and W. Tong, "Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method," *BMC Bioinformatics*, vol. 6, 2005.

[21] [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[22] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[23] [Online]. Available: http://www.cise.ufl.edu/d̃dd/cap6635/Fall-97/Short-papers/2.htm

[24] J. R. Quinlan, "C4.5: Programs for machine learning," *Machine Learning*, vol. 16, no. 3, pp. 235–240, 1994.

[25] [Online]. Available: http://www.cis.temple.edu/ ingargio/cis587/readings/id3-c45.html

[26] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 124–133, 1999.

[27] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[28] D. G. Altman, *Practical Statistics for Medical Research.*, Chapman and Hall, Eds. CRC Press, 1991.

[29] S. Hennessy, W. B. Bilker, J. A. Berlin, and B. L. Stromu, "Factors influencing the optimal control-to-case ratio in matched case-control studies," *American Journal of Epidemiology*, vol. 149, no. 1, pp. 195–197, 1999.

[30] W. Dupont, "Power calculations for matched case-control studies." *Biometrics*, vol. 44, pp. 1157–1168, 1988.

[31] W. Kim, D. Gordon, J. Sebat, K. Q. Ye, and S. J. Finch, "Computing power and sample size for case-control association studies with copy number polymorphism: Application of mixture-based likelihood ratio test," *PLoS ONE*, vol. 3, p. 3475, Oct 2008.

[32] F. D. Santis, M. P. Pacifico, and V. Sambucini, "Optimal predictive sample size for case-control studies," *Journal Of The Royal Statistical Society Series C*, vol. 53, pp. 427–441, 2004.

[33] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, pp. 1651–1686, 1998.