PERSPECTIVE OF MACHINE LEARNING IN EPIDEMIOLOGICAL STUDIES

Kontic B., Dzeroski S.

Jozef Stefan Institute, Ljubljana, Slovenia

ABSTRACT

A central problem in epidemiological studies is to find and verify associations between exposure and health consequences, based on field data. To alleviate the problem of manual searching for such associations, we explore the possibility of using machine learning to analyse epidemiological data.

Machine learning and knowledge discovery are concerned with extracting interesting and useful patterns, information, or knowledge from data. The advantages of machine learning methods over classical statistical methods include a richer and more structured form of knowledge discovered, where symbolic and numerical information are combined, and the ability to take into account existing domain knowledge. Machine learning methods have been successfully used in various domains of medical diagnosis and prognosis, as well as a variety of environmental domains. This study uses machine learning to find links between environmental and social factors, on one side, and health consequences, on the other.

Two approaches from the area of machine learning, i.e., decision tree induction and rule induction were used to study the influence of house heating practices, geographical location, smoking habits and some other variables on the occurrence of acute respiratory diseases. Diagnoses J00 to J20 according to the ICD-10 international classification were considered. A database of 760 patient records including the above variables was kindly provided by the Public Health Institute, Trnava, Slovak Republic.

The main goal of our study was to investigate the applicability of machine learning methods to epidemiological problems. Initial results indicate a positive answer: main influences to health effects (acute respiratory diseases) are successfully identified by machine learning methods.

METHODOLOGY AND TOOLS

We used several types of artificial intelligence (AI) tools in our experiment. Due to the length of this paper we can not describe the methods and tools in detail. The reader is kindly advised to consult the cited literature if more information is needed (1,2,3,4,5,6,7,8,9,10,11). The focus should be on decision tree-, and rule-induction systems.

In general, induction of decision trees is a classification problem, where AI tools can be seen as expert systems capable to automatically extract or construct knowledge from given instances of examples, described with independent variables (attributes) and dependent variable (the class). In our case, diagnoses J00 to J20 according to the ICD-10 international classification were the classes, all other variables were attributes. In Fig.1, a simplified decision tree is presented which shows classification of nonexposed and exposed children into different classes (diagnostic groups), e.g. J02, J03, J04, J06,.., according to the attributes age, sex, heating practices, etc.

Rules describe the influence of the attributes on selected class. In Table 1, a randomly selected rule from the rule list which was created in the experiment is presented. The presented rule says: if a child is more than 4.5 years old, and breastfeeding lasted less than seven months, and mother didn't smoke in the past, than in September we can expect diagnosis J03 for this child; numbers in brackets mean that there were four such examples in the dataset.

Table 2 explains the abbreviations of the attributes.

Problem definition

The subject of the experiment was investigation of relations between assumed influencing environmental and social parameters on occurrence of acute respiratory diseases in children living in selected areas of Slovak Republic. The machine learning system used for this experiment was C4.5, see Quinlan, 1993.

The test was performed on a database consisting of 760 examples. The influencing parameters were gender, age, residence at present, residence in past, breastfeeding, duration of breastfeeding, smoking of mother during pregnancy, ...altogether 33 attributes (see Table 2). The diagnoses J00 to J20 (ICD - 10) were studied health consequences.

Results

Several experiments were made under the study. The first and most general one was investigation of the influence of all attributes on selected diagnoses. The strongest influence was established by geographic location and the heating season (calendar month). This was more or less anticipated so in the next step of the experiment the attribute "geographic location" was ignored to see what is the residual dependency. Heating practices and age were attributes which gain importance.

The succeeding steps of the test consisted of checking the influence of different smaller sets of attributes on selected diagnoses (diseases). The first set of attributes was: gender, age, breastfeeding, duration of breastfeeding, smoking in house, cooking practices, heating practices, duration of gas heating in years, ventilation of fireplaces, visits of kindergarten, education of mother, education of father, type of house and exposition (altogether 13 attributes). The second set was: gender, age, breastfeeding, duration of breastfeeding, smoking of mother now and in the past, smoking of father now and in the past, cooking practices, heating practices, duration of gas heating in years, ventilation of fireplaces, visits of kindergarten, education of mother, education of father and type of house (altogether 15 attributes). The difference between these two sets is that "smoking in the house" was substituted with "smoking of mother" and "smoking of father" and that exposition of children was ignored in the second set. With such approach we wanted to learn which are the basic reasons (attributes) of influencing the stated diagnoses. Results show the strongest influence of the following attributes: exposure, age, cooking and heating practices.

```
Exp = nonexposed: J06 (443.8/279.0)
Exp = exposed:
|  Age <= 4.75 : J06 (137.5/86.8)
|  Age > 4.75 :
|  |  Cook = coal: J20 (5.1/4.5)
|  |  Cook = gas:
|  |  |  Heat = gas: J02 (6.1/5.5)
|  |  |  Heat = electricity: J03 (3.0/2.9)
|  |  |  Heat = coal: J03 (7.1/5.9)
|  |  |  Heat = wood:
|  |  |  |  G = male: J04 (5.5/4.3)
|  |  |  |  G = female: J11 (9.1/7.2)
|  |  Cook = electricity:
|  |  |  Age <= 5 : J20 (19.8/15.0)
|  |  |  Age > 5 :
|  |  |  |  G = male:
|  |  |  |  |  Heat = gas: J20 (1.4/1.1)
|  |  |  |  |  Heat = electricity: J20 (17.2/13.0)
|  |  |  |  |  Heat = coal: J03 (5.1/4.5)
|  |  |  |  |  Heat = wood: J11 (29.0/20.7)
|  |  |  |  G = female:
|  |  |  |  |  DurBF <= 1 : J11 (11.8/9.0)
|  |  |  |  |  DurBF > 1 : J03 (40.3/32.0)
|  |  Cook = wood:
|  |  |  G = male: J06 (9.1/6.4)
|  |  |  G = female: J20 (9.1/7.9)
```

Figure 1: Simplified decision tree: dependence of occurrence of selected acute respiratory diseases upon selected attributes

IF Age > 4.500000000
AND DurBF < 7.000000000
AND SmokeMPast = no
AND Month = september
THEN Diag = J03 [0 0 0 4 0 0 0 0 0 0 0 0]

Table 1: Randomly selected rule from the rule list

BN - birth numbers
G - gender
Age - age
RN - residence at present (name of village or city)
RP - residence in past
BF- breastfeeding (Y/N)
DurBF - duration of breastfeeding in month
SmokeMPreg - smoking of mother during

HeatingSource- source of heating in house or supply
Ventil - ventilation of fireplaces (outside - inside)
NofS - number of siblings
NofPeopleHousehold - number of people in household
NofR - number of rooms
ChildRoom - child-room (Y/N)
NofPeopleRoom - number of persons

RR = 0.35 (surprising, but explainable by social factors and different response rates), 95% CI = 0.29<RR<0.43, chi-square = 121.91, and P-value < 0.0000000.

CONCLUSION

The machine learning methods look very promising when using to investigate associations in the framework of epidemiological studies. They show consistency with classical statistical methods usually applied to show strength of these associations.
The study which was performed and presented in this paper is the basic level of potential use of these methods. Special utility of machine learning methods is, by our opinion, in the first stages of investigation of large datasets.

ACKNOWLEDGEMENT

We highly acknowledge understanding and help of dr.Gabriel Gulis from the Public Health Institute in Trnava, who kindly provided the necessary database to perform the study. Without his co-operation the study would not have been completed.

REFERENCES

1. Bratko I.: Machine Learning, In K.J.Gilhooly, editor, Human and Machine Problem Solving, Plenum Press, New York, 1989, p.265-287
2. Bratko I., Kononenko I.: Learning diagnostic rules from incomplete and noisy data. A1 methods in Statistics, Proceedings of the UNICOM seminar, London, 1986 (Also in Interactions in A1 and Statistics, B.Phelps, Ed., Gower Technical Press, 1987)
3. Quinlan J.R., Compton P., Horn K.A., Lazarus L.: Inductive knowledge acquisition: A case study, In J.R.Quinlan (Ed.), Applications of expert systems, Reading, MA: Addison-Wesley, 1987
4. Michalsky R.S., Chilausky R.L.: Learning by being told and learning from examples: An experimental comparison of the two methods for knowledge acquisition in the context of developing an expert system for soybean disease diagnosis, International Journal for Policy Analysis and Information Systems, 4 (2), 1980, p.125-161
5. Lavrac N., Varsek A., Gams M., Kononenko I., Bratko I.: Automatic construction of a knowledge base for a steel classification expert system, Proceedings of the 6th International Workshop on Expert Systems and their Application, Avignon, 1986
6. Mitchell T.M., Utgoff P.E., Banerji R.: Learning by experimentation: Acquiring and refining problem-solving heuristics, In R.S.Michalski, J.G.Carbonell & T.M.Mitchell (Eds.), Machine learning: An artificial intelligence approach, Palo Alto, Tioga, 1983
7. Winston P.H.: Learning structural descriptions from examples, In P.H.Winston (Ed.), The psychology of computer vision, McGraw-Hill, New York, 1975
8. Quinlan J.R.: Induction of decision trees, *Machine learning, 1*, 1986, p.81-106
9. Cestnik B., Kononenko I., Bratko I.: ASSISTANT 86: A knowledge elicitation tool for sophisticated users, In I.Bratko & N.Lavrac (Eds.), Progress in Machine Learning, Sigma Press, Wilmslow, 1987
10. Karalic A.: Employing linear regression in regression tree leaves, Proc. Of ECAI 92, Vienna, Austria, 1992
11. Quinlan J.R.: C4.5.Programs for machine learning, Morgan Kaufman, 1993