

# Observational Logic Integrates Data Mining Based on Statistics and Neural Networks

Martin Holeňa

Institute of Computer Science, Academy of Sciences of the Czech Republic  
Pod vodárenskou věží 2, CZ-182 00 Praha 8, Czech Republic,  
[martin@cs.cas.cz](mailto:martin@cs.cas.cz), <http://www.cs.cas.cz/~martin>

**Abstract.** In data mining, artificial neural networks have become one of important competitors of traditional statistical methods. They increase the potential of discovering useful knowledge in data, but only if the differences between both kinds of methods are well understood. Therefore, integrative frameworks are urgently needed. In this paper, a framework based on the calculus of observational logic is presented. Basic concepts of that framework are outlined, and it is explained how generalized quantifiers can be defined in an observational calculus to capture data mining with statistical and ANN-based methods.

## 1 Introduction

In the area of data mining, traditional statistical methods are often competed by methods relying on more recent approaches, including artificial neural networks, [3, 5, 15]. Those methods increase the potential of knowledge discovery and can bring us closer to the aim of extracting all the useful knowledge contained in the data. However, a prerequisite is that specific features of each kind of methods (e.g., the ability of statistical methods to discover structures which cannot be attributed to random influences and noise, or the universal approximation capabilities of neural networks), the differences between them with respect to the applicability conditions and to the meaning of the obtained results are well understood. Therefore, integrative frameworks are urgently needed.

In this paper, an integrative framework based on the calculus of observational logic is presented. It has been elaborated already in the seventies, as a theoretical foundation of the method Guha, one of the first methods for automated knowledge discovery [7]. Originally, observational calculus has been proposed as a means for a unified treatment of different kinds of statistical methods. Recent research indicates that the framework can be successfully extended also behind statistical methods of knowledge discovery [8, 12, 14]. Basic concepts pertaining to the calculus are recalled in the following section, whereas the remaining two sections explain how generalized quantifiers can be defined in the calculus to capture statistical hypotheses testing and the extraction of rules from data with artificial neural networks. The framework is illustrated on small fragments from two real-world applications.

## 2 Observational Logic and Its Importance for Data Mining

*Monadic observational predicate calculus* is a collection of *unary predicates* with a single object variable  $x$ , and of *generalized quantifiers*. A unary predicate, in general, states a range of values or even one specific value of a variable capturing some single property of an individual object (e.g., a patient), such as in

disease duration > 10 years,  
sex = male.

Consequently, an open formula built from unary predicates by means of logical connectives corresponds to some combined property (combination of single properties), of an individual object, e.g.,

sex = male  $\wedge$  disease duration > 10 years  $\wedge$  grand mal seizures.

On the other hand, closed formulae state properties characterizing in some way the whole set of objects (e.g., patients) underlying the considered data. For example,  $(\forall x) \text{ age} > 15$ , meaning that we deal with data about adult patients. In the case of binary quantifiers or other quantifiers of a higher arity, the closed formula  $(Qx)(\varphi_1, \dots, \varphi_m)$ , built from an  $m$ -ary generalized quantifier  $Q$  and open formulae  $\varphi_1, \dots, \varphi_m$ , in general states some relationship between properties corresponding to  $\varphi_1, \dots, \varphi_m$ . Examples will be given in the next section.

As *models* of a monadic observational predicate calculus, *data matrices* are used, which for the considered objects record the values of variables capturing the considered properties. More precisely, each column records the values of one variable. The interpretation in the model of a predicate concerning that variable is a  $\{0, 1\}$ -valued column vector with 1s on positions corresponding to those rows of the data matrix for which the values of the variable fulfil the predicate. Similarly, the interpretation of an open formula is the column vector that results from combining the interpretations of its constituent predicates in the boolean algebra of  $\{0, 1\}$ -valued vectors. For example, the interpretation of

sex = male  $\wedge$  disease duration > 10 years  $\wedge$  grand mal seizures,

is the vector with ones on positions corresponding to those rows of the data matrix for which the three constituent predicates are simultaneously valid, and with zeros on all remaining positions. Finally, the interpretation of the closed formula  $(Qx)(\varphi_1, \dots, \varphi_m)$  is obtained by applying some  $\{0, 1\}$ -valued function  $\text{Tf}_{(Qx)(\varphi_1, \dots, \varphi_m)}$ , called the *truth function* of  $(Qx)(\varphi_1, \dots, \varphi_m)$  to the data matrix serving as the model. The value of the truth function can, in general, depend on the whole data matrix, more precisely on those its columns that are needed to obtain the interpretations of  $\varphi_1, \dots, \varphi_m$ . However, in many important cases that dependence reduces to a dependence on the interpretation of  $\varphi_1, \dots, \varphi_m$ . In such a case, the truth function is a function on the set of all dichotomous matrices with  $m$  columns, is denoted simply  $\text{Tf}_Q$  and called the truth function of the quantifier  $Q$ . For example, the truth functions of the classical unary *existential quantifier*  $\exists$  and *universal quantifier*  $\forall$  are defined  $\text{Tf}_\exists(M) = 1$  iff  $\sum M_i > 0$  and  $\text{Tf}_\forall(M) = 1$  iff  $M_1 = M_2 = \dots = 1$ , respectively.

If the observational calculus includes at least one of the quantifiers  $\forall$  and  $\exists$ , all tautologies derivable in the classical boolean predicate logic can be derived also here. In addition, a number of other tautologies are derivable also for various classes of generalized quantifiers of a higher arity [7]. Using those tautologies, the validity of some relationships can be deduced from the validity of other relationships, without having to evaluate the formula capturing the deduced relationship. This can substantially increase the efficiency of data mining.

## 3 Generalized Quantifiers for Statistical Hypotheses Testing

Let the data matrix serving as a model of a monadic observational calculus can be viewed as a realization of a random sample from some probability distribution  $P$ .

Through appropriate constraints on  $P$ , it is possible to capture probabilistic relationships between the properties corresponding to the columns of that data matrix. Assuming that  $P$  is a priori known to belong to some set  $\mathcal{D}$  of probability distributions, a *statistical test* assesses the validity of the constraint  $H_0 : P \in \mathcal{D}_0$  for a given  $\mathcal{D}_0 \subset \mathcal{D}$  against the alternative  $H_1 : P \in \mathcal{D}_1 = \mathcal{D} \setminus \mathcal{D}_0$ . The test is performed using some random variable  $t$ , the test statistic, and some borel set  $C_\alpha$ , the critical region of the test on a significance level  $\alpha \in (0, 1)$ , which are connected to  $H_0$  through the condition  $(\forall P \in \mathcal{D}_0) P(t \in C_\alpha) \leq \alpha$ . In addition, the test often makes use of the fact that if the considered model is a realization of a random sample, then also the matrix  $M$  of interpretations of open formulae  $\varphi_1, \dots, \varphi_m$  is a realization of a random sample, more precisely a realization of a random sample from a multinomial distribution. That fact allows to choose some test statistic  $t$  that is a function of that multinomial random sample, and to define a generalized quantifier  $\sim$  corresponding to the constraint  $H_1$  by means of a truth function depending only on  $M$ ,

$$\text{Tf}_{\sim}(M) = 1 \text{ iff } t_M \in C_\alpha, \quad (1)$$

where  $t_M$  is the realization of  $t$  provided the realization of the random sample from  $P$  yields the matrix of interpretations  $M$ .

The existing implementations of Guha include about a dozen quantifiers based on statistical hypotheses testing [7, 16]. All of them are binary, the most fundamental being the *lower critical implication*  $\rightarrow_\alpha^p$ ,  $p \in (0, 1)$ , which corresponds to the binomial test (with a significance level  $\alpha \in (0, 1)$ ), the *Fisher quantifier*  $\sim_\alpha^F$ , corresponding to the one-sided Fisher exact test, and the  $\chi^2$  *quantifier*  $\sim_\alpha^{\chi^2}$ , corresponding to the  $\chi^2$  test. For example in [21], the following relationships have been discovered using the Fisher quantifier:

```
sex = male  $\wedge$  disease duration > 10 years  $\wedge$ 
 $\wedge$  only grand mal seizures  $\sim^F$  memory quotient > 90,
disease duration > 10 years  $\wedge$  only grand mal seizures  $\wedge$ 
 $\wedge$  course of the disease = good  $\sim^F$  memory quotient > 90.
```

A second example comes from an application of Guha to the area of river ecology [13]. Table 1 shows antecedents of formulae found to hold in the collected

data and expressible as  $(\bigwedge_{i \in \mathcal{I}} E_i \sim_{10\%}^F \text{RD}) \wedge (\bigwedge_{i \in \mathcal{I}} E_i \rightarrow_{10\%}^{\frac{2}{3}} \text{RD})$ , where  $\mathcal{I}$  is some subset of ecological factors,  $E_i$  for  $i \in \mathcal{I}$  are predicates stating ranges of values of the respective ecological factors, and  $\text{RD}$  is a predicate stating the occurrence of the species “Robackia demeierei”. More precisely,  $\text{RD}$  states that the number of individuals of “Robackia demeierei” per sample is at least  $\frac{1}{10}n_{\max}$ , where  $n_{\max}$  denotes the maximal number of individuals of this species per sample encountered in the data.

#### 4 Quantifiers for ANN-Based Rule Extraction

The extraction of knowledge from data by means of artificial neural networks has received much attention not only in connection with data mining, but also in connection with pattern recognition [1, 2, 17–19]. Actually, already the mapping learned by a neural network incorporates knowledge about the implications that certain values of the input variables have for the values of the output variables. Usually, however, the ANN-based knowledge extraction aims at the more comprehensible representation of those implications as *rules* [4, 10, 18, 20].

A comprehensive survey of ANN-based rule extraction methods has been given in [1] (see also [19]), where a classification scheme for those methods has been proposed as well. In the context of that classification scheme, two remarks

**Table 1.** Antecedents of found formulae  $(\bigwedge_{i \in \mathcal{I}} E_i \sim_{10\%}^F \text{RD}) \wedge (\bigwedge_{i \in \mathcal{I}} E_i \rightarrow_{10\%}^{\frac{2}{3}} \text{RD})$ 

attained significance:	$\sim_{10\%}^F$	$\rightarrow_{10\%}^{\frac{2}{3}}$
90%il of grain diameter < 3mm $\wedge$ flow velocity > 0.2m/s	0.008 %	3.6 %
mean grain diameter < 1.5 mm $\wedge$ flow velocity > 0.2m/s	0.008 %	3.6 %
glowable proportion < 5% $\wedge$		
90%il of grain diameter < 3mm $\wedge$ flow velocity > 0.2m/s	0.33 %	5.5 %
glowable proportion < 5% $\wedge$		
mean grain diameter < 1.5mm $\wedge$ flow velocity > 0.2m/s	0.33 %	5.5 %

can be made to the possibility to capture the extraction of rules using a generalized quantifier of an observational calculus:

- As far as the primary dimension of the proposed scheme is concerned, the *expressive power* of the rules, the fact that the truth function of a generalized quantifier is  $\{0, 1\}$ -valued implies that only *boolean rules* can be covered. In connection with the other major class of rules, fuzzy rules, ongoing research into fuzzy generalized quantifier should be mentioned [6, 9, 11, 12].
- With respect to the second dimension, the *translucency*, the extracted rules should concern only network inputs and outputs. The reason is that hidden neurons are not explicitly assigned to properties of real objects, thus they do not correspond to open formulae of an observational calculus.

Even with the choices with respect to those two dimensions fixed, there are innumerable possibilities how to design a rule extraction method. Here, a particular method described in [14] is considered. Due to the limited extent of this contribution, the method cannot be recalled in detail, instead only its basic principles are listed:

1. As the neural network architecture, a *multilayer perceptron (MLP)* with one hidden layer,  $m$  input neurons and  $n$  output neurons is used.
2. To each hidden and output neuron, a *piecewise-linear sigmoidal activation function* is assigned.
3. Any hyperrectangle in  $\mathbb{R}^n$  (in particular, any set of output value tuples for which a given conjunction of unary predicates holds) is the map, via  $F$ , of some finite set of polyhedra in  $\mathbb{R}^n$ .
4. A replaceability relation  $\approx_{\varepsilon}^{\mu_P}$  is defined between the set  $\mathcal{P}_m$  of all polyhedra in  $\mathbb{R}^n$  and the set  $\mathcal{H}_m$  of all hyperrectangles in  $\mathbb{R}^n$ , being determined by as system of monotone measures  $\mu = (\mu_P)_{P \in \mathcal{P}_m}$ , and by a tolerance  $\varepsilon > 0$ . Practically important monotone measures are the empirical distribution of the data  $\hat{\mu}$ , and its weighted version  $\hat{\mu}_P$ , which is for  $P \in \mathcal{P}_m$  with  $\hat{\mu}(P) \neq 0$  defined  $(\forall S \subset \mathbb{R}^m) \hat{\mu}_P(S) = \frac{\hat{\mu}(S)}{\hat{\mu}(P)}$ . For  $P \in \mathcal{P}_m$  with  $\hat{\mu}(P) = 0$ , the measure  $\hat{\mu}_P$  can be defined arbitrarily, depending on whether we want to admit replaceability of such polyhedra.

This method allows to state the fact that particular values of some input variables imply particular values of some output variables, using the formula

$$\bigwedge_{k \in \mathcal{I}} \pi_k \rightarrow_{\varepsilon}^{\mu} \bigwedge_{k \in \mathcal{O}} \pi_k \quad (2)$$

(more pedantically,  $(\rightarrow_{\varepsilon}^{\mu} x)(\bigwedge_{k \in \mathcal{I}} \pi_k, \bigwedge_{k \in \mathcal{O}} \pi_k)$ ), where  $\mathcal{I} \subset \{1, \dots, m\}$ ,  $\mathcal{O} \subset \{1, \dots, n\}$ , for each  $k \in \mathcal{I}$ , or  $k \in \mathcal{O}$ ,  $\pi_k$  is a unary predicate stating some

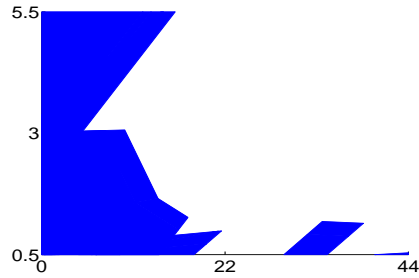


Fig. 1. Example of found polyhedra

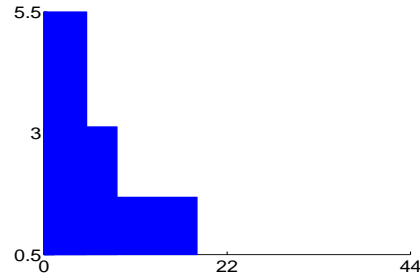


Fig. 2. Example of found hyperrectangles

interval  $I_k$  of values of the  $k$ -th input variable, or the  $k$ -th output variable, respectively, and  $\rightarrow_\epsilon^\mu$  is a binary quantifier corresponding to the replaceability relation  $\approx_\epsilon^{\mu_P}$ . The reader is defined to [14] for the definition of its truth function.

For comparison with Table 1, Figure 1 depicts a 2-dimensional projection of the union of polyhedra that a particular coordinate  $F_k$  of the mapping  $F$  learned by a MLP maps to a particular interval  $I_k$ . In Figure 1, the considered  $F_k$  counts the number of individuals of “Robackia demeierei” per sample, and  $I_k = \langle \frac{1}{10}n_{\max}, +\infty \rangle$ . The depicted union of polyhedra is projected to the coordinates corresponding to the ecological factors “glowable proportion” and “90%il of grain diameter”. Figure 2 shows a projection, to the same coordinates, of those hyperrectangles that replace any polyhedron  $P$  from the union depicted in Figure 1 according to the replaceability relation  $\approx_{10\%}^{\hat{\mu}_P}$ , where  $\hat{\mu}_P$  is the weighted empirical distribution. The projection in Figure 2 reflects the following rules that can be extracted from the data using (2) with  $\mu = (\hat{\mu}_P)_{P \in \mathcal{P}_m}$ :

```
glowable proportion < 5.39 %  $\rightarrow_{10\%}^\mu$  RD,
glowable proportion < 18.69 %  $\wedge$  mean grain diameter < 1.17 mm
 $\wedge$  90%il of grain diameter < 1.7 mm  $\rightarrow_{10\%}^\mu$  RD,
glowable proportion < 9.04 %  $\wedge$  mean grain diameter < 1.62 mm
 $\wedge$  90%il of grain diameter < 3.06 mm  $\rightarrow_{10\%}^\mu$  RD.
```

Needless to say, it is now very easy to compare such extracted rules with result obtained using statistical hypotheses testing. This impressively illustrates the usefulness of the employed integrative framework.

## 5 Conclusion

This paper tried to show that observational logic can be used as a framework for integrating traditional statistical methods for knowledge discovery in data with methods based on artificial neural networks. This framework provides a common theoretical view of both kinds of methods while preserving their specific advantages. Due to the different underlying paradigms and different initial assumptions, artificial neural networks do not necessarily yield the same results as statistical methods. Hence, a coincidence between relationships discovered in the data by both kinds of methods increases the chance that those relationships pertain to the reality behind the data, to the phenomena that generated them. On the other hand, if a relationship found by means of some quantifier can not be confirmed using other quantifiers, including a quantifier based on the other paradigm, then such as relationship is a natural starting point for further, deeper investigations.

## Acknowledgement

The research reported in this paper has been supported by the grant 201/00/1489 of the Grant Agency of the Czech Republic.

## References

1. R. Andrews, J. Diederich, and A.B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge Based Systems*, 8:378–389, 1995.
2. C.M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon-Press, Oxford, 1995.
3. K. Cios, W. Pedrycz, and R. Swiniawski. *Data Mining Methods for Knowledge Discovery*. Kluwer Academic Publishers, Dordrecht, 1998.
4. W. Duch, R. Adamczak, and K. Grabczewski. Extraction of logical rules from neural networks. *Neural Processing Letters*, 7:211–219, 1998.
5. Y. Frayman and L. Wang. Data mining using dynamically constructed recurrent fuzzy neural networks. In *Research and Development in Knowledge Discovery and Data Mining*, pages 122–131. Springer-Verlag, Berlin, 1998.
6. P. Hájek. *Metamathematics of Fuzzy Logic*. Kluwer Academic Publishers, Dordrecht, 1998.
7. P. Hájek and T. Havránek. *Mechanizing Hypothesis Formation*. Springer-Verlag, Berlin, 1978.
8. P. Hájek and M. Holeňa. Formal logics of discovery and hypothesis formation by machine. In *Discovery Science*, pages 291–302. Springer-Verlag, Tokyo, 1998.
9. P. Hájek and L. Kohout. Fuzzy implications and generalized quantifiers. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 4:225–233, 1996.
10. M.J. Healy and T.P. Caudell. Acquiring rule sets as a product of learning in a logical neural architecture. *IEEE Transactions on Neural Networks*, 8:461–474, 1997.
11. M. Holeňa. Exploratory data processing using a fuzzy generalization of the Guha approach. In *Fuzzy Logic*, pages 213–229. John Wiley and Sons, New York, 1996.
12. M. Holeňa. Fuzzy hypotheses for Guha implications. *Fuzzy Sets and Systems*, 98:101–125, 1998.
13. M. Holeňa. Traditional and modern artificial intelligence explores ecological data. In *STeP 2000: Millenium of Artificial Intelligence*, 2000.
14. M. Holeňa. Statistical, logic-based, and neural networks based methods for mining rules from data. In *Multisensor and Sensor Data Fusion*. NATO Science Series Publishers, to appear.
15. H. Lu, R. Setiono, and H. Liu. Effective data mining using neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 8:957–961, 1996.
16. J. Rauch. Logical calculi for knowledge discovery in databases. In *Principles of Data Mining and Knowledge Discovery*, pages 47–57. Springer-Verlag, Berlin, 1997.
17. B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
18. R. Setiono. Extracting rules from neural networks by pruning and hidden unit splitting. *Neural Computation*, 9:205–225, 1997.
19. A.B. Tickle, R. Andrews, M. Golea, and J. Diederich. The truth is there: Directions and challenges in extracting rules from trained artificial neural networks. *IEEE Transactions on Neural Networks*, 9:1058–1068, 1998.
20. G.G. Towell and J.W. Shavlik. Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 13:71–101, 1993.
21. J. Zvárová, J. Preiss, and A. Sochorová. Analysis of data about epileptic patients using Guha method. *International Journal of Medical Informatics*, 45:59–64, 1997.