

# Determining property relevance in concept formation by computing correlation between properties

*João José Furtado Vasco*

*UNIFOR – Universidade de Fortaleza*

*Centro de Ciências Tecnológicas – Departamento de Computação*

*Av. Washington Soares 1321*

*Fortaleza - CE Brazil*

*[vasco@ufc.br](mailto:vasco@ufc.br)*

**Abstract.** We propose a method to incrementally compute and use, during the concept formation, the relevance of a concept property. This relevance is computed through the account of the property correlation with other ones and it is used by the concept quality function in order to improve predictive accuracy. The proposed approach is analyzed concerning both the prediction power of the generated concepts and the time and space complexity of the concept formation algorithm.

## 1. Introduction

The general aim of concept formation is to construct, based on entity descriptions (observations), a (usually hierarchical) categorization of entities. Each category is provided with a definition, called a concept, which summarizes its elements. Further aim is to use concepts to categorize new entities and to make predictions concerning unknown values of attributes of these entities. Therefore, quality of a concept can be measured in terms of its ability to make *good* predictions about unknown values of attributes (the prediction power). Unlike supervised systems, in which concept quality is measured from the capacity in discovering a value for a single property, in concept formation, the quality of a concept is measured by its capacity in allowing prediction of values for several attributes.

An important aspect that must be considered in concept formation concerns the relevance (sometimes called salience) of particular attributes/values (or properties). Cognitive psychology [Seifert 89] and machine learning [Stepp 86], [Decaestecker 93] researchers have pointed out the necessity to determine how much a certain property is relevant within a concept.

In this paper, we propose a method to compute the relevance of a concept property based on the correlation between properties of the entities covered by the concept. We describe our approach using a COBWEB-based algorithm, called FORMVIEW, which can generate several hierarchies of categories describing different perspectives [Vasco 97]. In a multi-perspective context, the relevance of a property is crucial because it determines the hierarchical organization of categories. Since FORMVIEW uses a probabilistic representation, correlation between properties is computed from conditional probabilities. However, the proposed approach is generic and can be employed in algorithms, which may use other representations.

## 2. Concept Formation Systems

Concept formation (CF, for short) [Fisher 87] recognize regularities among a set of non-preclassified entities and induce a concept hierarchy that summarizes these entities. A CF algorithm is reduced to a search, in the space of the all possible concept hierarchies, for that one that covers the observed entities and optimizes an evaluation function measuring a quality criterion. In concept formation entities are treated one after another as soon as they are observed and the classification of new entities is made by their adequacy for the existing conceptual categories.

Typically concept representation is probabilistic, in which concepts have a set of attributes and all possible values for them. Each concept has the probability that an observation is classified into the concept and each value of a concept attribute has associated a *predictability* and a *predictiveness* [Fisher 87]. The predictability is the conditional probability that an observation  $x$  has value  $v$  for an attribute  $a$ , given that  $x$  is a member of a category  $C$ , or  $P(a=v|C)$ . The predictiveness is the conditional probability that  $x$  is member of  $C$  given that  $x$  has value  $v$  for  $a$  or  $P(C|a=v)$ .

## 3. Concept Formation taking into account the Relevance of a Property

### 3.1 Using the relevance of a property in concept formation

FORMVIEW's category quality (UCF) is defined as the increase in the expected number of properties that can be correctly predicted given knowledge of a category over the expected number of correct predictions without such knowledge. FORMVIEW, in addition, takes into account the relevance of the properties. We consider that the increase in the expected number of properties to be predicted from a category depends on the relevance of its properties. Formally, UCF is :

$$UCF(C) = \left( \sum_{i=1}^p \Delta(p_i) P(p_i|C) P(C|p_i) - P(C) P(p_i)^2 \right) \quad (1)$$

Where  $\Delta(p_i) = (\text{the relevance of the category property } p_i + P(p_i))$ .

To compute a property relevance, FORMVIEW uses a strategy that relies on attribute dependence (or attribute correlation) in the way that was defined by Fisher [Fisher 87]. Formally, the dependence of an attribute  $A_m$  on other  $n$  attributes  $A_i$  can be defined as :

$$Mdep(A_m, A_i) = \frac{\left( \sum_i^n \sum_{j_i} P(A_i = V_{j_i}) \sum_{j_m} [P(A_m = V_{j_m} | A_i = V_{j_i})^2 - P(A_m = V_{j_m})^2] \right)'}{n} \quad (2)$$

Where  $V_{j_i}$  signifies the  $j_i$ th value of attribute  $A_i$  and  $A_i \neq A_m$ .

In fact, this function measures the average increase in the ability to guess a value of  $A_m$  given the value of a second attribute  $A_i$ . We consider that this strategy

accounts for the relationship between attribute dependence and the ability to correctly infer an attribute's value using a probabilistic concept hierarchy. We can thus determine those attributes that depend on others and, as a consequence, those that influence the prediction of others. By an *influent attribute*, we mean that, if we know its value, it allows a *good* prediction about the value of others. We have thus defined the total influence  $Tinfl$  of an attribute  $A_k$  on others  $A_m, m=1, \dots, n$ , as the following:

$$Tinfl(A_k) = \frac{\sum_{m=1}^n Dep(A_m, A_k)}{n} \quad \text{where } A_k \neq A_m. \quad (3)$$

Our claim is that attribute dependence gives a measure to ponder attribute relevance, which, in this context, means how much an attribute correlates with others.

### 3.2 Computing property relevance for each concept

In formula 2, we have defined the probability of predicting a property  $p$  given another property  $p'$ ;  $P(p|p')$ . Actually, for each concept  $C$  within a hierarchy, we have  $P(p|p'$  and  $C$ ). The acquisition of this probability is problematic, since it cannot be computed only from the *predictability* and *predictiveness* stored by FORMVIEW. Instead of keeping all the 2x2-property correlation, which would take too much space, or of computing such a correlation for each new observation, which would be computationally expensive, we have defined a procedure that implements a tradeoff between time and space requirements.

Our procedure consists of maintaining two triangular arrays which keep the 2x2-property correlation:  $T-root$  and  $T-son$ . These arrays keep such a correlation for all the observations already seen.  $T-root$  is updated once for each new observation. It allows computing the relevance of the root's properties.  $T-son$  accompanies side by side the path followed by the observation during its categorization. It is updated at each hierarchical level until the observation arrives at the leaves.

Two procedures are responsible for updating the frequency of a property correlation:  $UpdateArray$  and  $RefineTson$ . In  $UpdateArray$ , the frequency of a property correlation is computed for all the concept properties. The procedure  $RefineTson$  refines  $T-son$  each time FORMVIEM descends the concept hierarchy. This refinement consists of updating  $T-son$  in order to let it only with the account of the existing correlation between the properties of observations covered by the current root concept. For each concept,  $T-son$ 's actualization is based on the following strategy: if the quantity of observations covered by a concept is greater than the total of its *brothers* (children of the concept's father),  $T-son$  is updated from those observations which are covered by these later. Otherwise,  $T-son$  only stores the property correlation from observations covered by a concept.

Finally, the procedure  $ComputeRelevance$  computes the relevance of a property using the conditional probability that an entity has a property given that another property is known. These properties are computed from the frequency of a property correlation represented in  $T-son$ .

## 4. Performance Task

In our process of evaluation of FORMVIEW, we pay attention to the prediction power of a hierarchy generated by it. The basic idea is to submit a set of «questions» (normally, incomplete observations) to the system, whose answer is based on the generated representation. The quality of the representation is measured according to its capacity to give «good» answers (i.e. to infer values for attributes).

We have used three test domains taken from UCI machine-learning dataset : two animal classification domains (ZOO domains) and the Pittsburgh's bridges domain. The results obtained in the tests were useful to show more clearly the contribution of the use of predictive influence as a heuristic to compute the relevance of a property. Indeed, hierarchies generated from this heuristic have a better prediction power than those generated by other systems. It is due to the fact that concepts are organized around properties having a strong predictive influence. Thus, when one infers the value of a property, he/she increases the probability to infer values for other properties influenced by the first one. Figure 1 and Figure 2 illustrate tests done in ZOO domains. The results of tests done in Pittsburgh domain can be viewed in [Vasco 97].

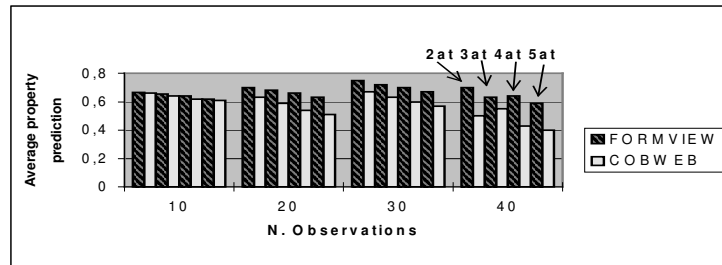


Fig. 1 Prediction of several attributes : ZOO Physiologic

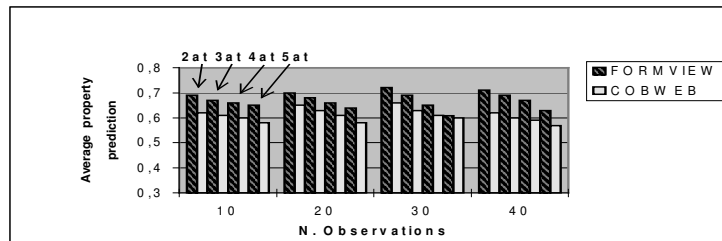


Fig. 2 Prediction of several attributes : ZOO Pet

## 5. Complexity

To examine the time and space complexity required by FORMVIEW's approach, we consider  $n$  the number of categorized entities and  $L$  the average branch factor of the concept tree. The cost of the procedure for computing property relevance is bound to the cost of updating  $T-root$  and  $T-son$ . It should be reminded that, FORMVIEW stores the  $2 \times 2$ -correlation for each concept property, in these arrays.

First, we determine the necessary space to stock these correlations. Let *cell* be the unit where the frequency of correlation between two properties will be kept. In each array that maintains the frequency of correlation between properties, for  $n$  observations, which have in average  $nbAT$ , attributes with in average  $nbV$  values, we need

$$\sum_{i=1}^{nbAT \times nbV} i \text{ cells, that is to say, } O(nbAT \times nbV)^2 \text{ cells.}$$

As for the time complexity, we have the cost to update *T-root* and *T-son*. when a new observation is categorized. The cost of *T-root*'s update is the same as that required in space. *T-son*'s updating is more expensive than that of *T-root* because it must be done to each level of the hierarchy (on average time  $\log_L^n$ ). For each level, only the frequencies of correlation between observation properties covered by the current node must be represented. Thus, *T-son* must be actualized  $m$  times ( $m < n$ ), where  $m$  represents the minimum between the number of observations which are not covered by the current node and the number of observations covered by the current node. In the worst case, we have  $m = n/2$ , which would make the geometric progression ( $n/2, n/4, n/8, \dots, 1$ ) for all the depth of the tree. The cost of *T-son* actualization is thus the order:  $O(2^{n/2} \times (nbAT \times nbV)^2)$ .

Finally, it is necessary to mention that the cost of computing the predictive influence for every concept also requires time effort of the order  $O(nbAT \times nbV)^2$ .

The total cost for computing the predictive influence of property is:

$$O(2^{n/2} \times (nbAT \times nbV)^2 + \log_L^n \times (nbAT \times nbV)^2) = O((nbAT \times nbV)^2 (2^{n/2} + \log_L^n))$$

This reflects how expensive, in time, our approach is, compared to COBWEB.

## 6. Related Work

The definition of a property relevance has been treated in early incremental concept formation systems. ADECLU [Decaestecker 93] uses a statistical measure to quantify the correlation between the property of a concept and the variable "membership of the concept". It maintains a 2x2-contingence table for each property of each concept. However, there is no account of the correlation between the properties. In ECOBWEB [Reich 91], property relevance is taken into account in the categorization process in the same way we have implemented here. However, the information of which properties are relevant is given by the user. Cluster/CA [Stapp 86] equally uses the information about property relevance defined by the user in the GDN (Goal Dependence Network). Early versions of FORMVIEW also follow this same idea [Vasco 96]. The non-incremental system WITT [Hanson 89] computes the correlation between properties to create categories. It keeps for each (concept, property) pair a contingence table. Such a table contains the frequency of simultaneous occurrence of property pairs. For  $A$  attributes, WITT keeps for each

concept  $A(A-1)/2$  contingency tables. This can be a very tough requirement in terms of space.

## 7. Conclusion and Future Research

We have defined a method to compute and use the relevance of a property in concept formation systems. The first results obtained with the use of property relevance are encouraging. They shown us that, FORMVIEW's approach can provide representations it generates with better prediction power than those generated from systems that do not take into account the relevance of properties. However, additional tests are necessary, mainly with regards to evaluate the performance of FORMVIEW with more data. Moreover, these tests will be useful to analyze, if the gains in accuracy compensates the computational costs required by the method. In order to analyze the generality of the proposed method, future research consists of the application of this method to systems which use different representations.

## References

- [Decaestecker 93] Decaestecker, C. : Apprentissage et outils statistiques en classification conceptuelle incrémentale. *Revue d'Intelligence Artificielle*, v 7, n. 1, 1993.
- [Fisher 87] Fisher, D.H.: *Knowledge Acquisition via Incremental Conceptual Learning*. Machine Learning, vol 2, numero 2, 1987.
- [Fisher 91] Fisher, D., Pazzani, M., Langley, P. : *Concept Formation: Knowledge and Experience in Unsupervised Learning*. Morgan Kaufmann, 1991.
- [Hanson 89] Hanson, S., Bauer, M. : Conceptual Clustering, Categorization, and Polymorphy. *Machine Learning*, v. 3, pp : 343-372, 1989.
- [Reich 91] Reich, Y., Fenves, S. : The Formation and Use of Abstract Concepts in Design. In [Fisher 91].
- [Seifert 89] Seifert, C.: *A Retrieval Model Using Feature Selection*. Proc. of the 6th Int. Workshop on ML. Morgan Kaufmann. 1989.
- [Stepp 86] Stepp, R., Michalski, R.: *Conceptual Clustering: Inventing goal-oriented classifications of structured objects*. In Michalski, R., Carbonnel, J., Mitchell, T. (Eds), *Machine Learning, An Intelligence Approach*. Vol II. Morgan Kaufmann, CA. 1986.
- [Vasco 96] Vasco, J.J.F., Faucher, C., Chouraqui, E.: *A Knowledge Acquisition Tool for Multi-perspective Concept Formation*. In proc. of European Knowledge Acquisition Workshop - EKAW-96. Shadbolt, Shreiber and O'Hara (Eds). Springer Verlag; 1996.
- [Vasco 97] Vasco, J.J.F., Formation de concepts dans le contexte des langages de schémas. Thèse de doctorat. Université d'Aix Marseille III, IUSPIM/DIAM, 1997.