NEW SIMPLE MEASURES FOR INDUCTIVE LEARNING

G. Ramos-Jiménez, R. Morales-Bueno, A. Villalba-Soria*
Dpt. Lenguajes y Ciencias de la Computación
E.T.S. Ingeniería Informática
Universidad de Málaga
Aptdo. 4114, 29080-Málaga (SPAIN)
Tlfno. 95-2132725, FAX: 95-2131397
e-mail: ramos@lcc.uma.es
* Student of E.T.S. Ingeniería Informática

ABSTRACT.- A family (*b*eta) of new attribute selection measures for TDIDT (Top Down Induction Decision Trees) algorithms is introduced. These measures are based upon the majority class (*b*est class). They are easy to compute and produce good empirical results.

KEYWORDS: Machine learning, Induction, Decision Tree, ID3, TDIDT.

1 INTRODUCTION

Inductive learning is a branch of Artificial Intelligence. In this branch the family of TDIDT algorithms is well known *Buntine (1992) Michalski (1983)*. From a set of experiences, these algorithms build a decision tree *Cuena (1987) Hunt (1996)*. ID3 *Quinlan (1979) (1986)* is the best known algorithm of this family. Developed by J.R. Quinlan in 1979, this algorithm builds decision trees by using the Entropy *Shannon (1948)* as measure. Other measures are the gain ratio *Quinlan (1986)* and the normalized distance *López de Mántaras (1991) (1992)*. In general, every TDIDT algorithm *Fayyad (1992) Quinlan (1993) Schlimmer (1986) Utgoff (1989)* needs a measure.

In this paper we define three new measures *b1*, *b2*, and *b3*. All measures are based upon the majority class (the *b*est class).

In section 2 we develop the previous concepts that we will use. In section 3 we define the new measures and it will be clear that they are easy to compute. The goodness of the measures is shown in section 4 by means of experimental results; then some conclusions are obtained.

2 NOTATIONS AND PREVIOUS CONCEPTS

We assume knowledge of TDIDT algorithms and the measures above cited. Only concepts necessary in order to define the new measures are introduced.

The following notation is according to our work *Ramos-Jiménez (1999)*:

A *problem* with a attributes and k class is a vector $(\underline{m}, k) \in \mathbb{N}^a \times \mathbb{N}$, where $\underline{m} = (m_1, ..., m_a)$.
The domain of each attribute X_i is called $D_i = \{1, 2, ..., m_i\}$, $i = 1, ..., a$.
Also, to avoid partial functions an attribute X_0 is included with domain $D_0 = \varnothing$.
The attribute set is denoted by $\mathbf{X}$: $\mathbf{X} = \{X_0, X_1, ..., X_a\}$.
An special attribute with k values called *class* is denoted by C , and his domain D is codified $D = \{1, 2, ..., k\}$.

Let us consider a problem $(\underline{m} , k) \in \mathbb{N}^a \times \mathbb{N}$, with a attributes whose domains are $D_1, D_2, \ldots , D_a$. We define the universe of experiences $U_E = D_1 \times D_2 \times \ldots \times D_a \times D$. An experience e is an element of U_E, that is, a vector with $a+1$ components: $e = (X_1(e), X_2(e), \ldots, X_a(e), C(e)) \in U_E$, where $X_i(e)$ is the value of the i attribute in the e experience, and $C(e)$ represents the last component $(a+1)$; so, $C(e)$ is the value of *class* in the e experience.
We will work with finite sequences of experiences $E = \{e_1 , e_2 , \ldots, e_N\}$, where some elements could be repeated.
The set of all finite sequences of experiences is represented by $\mathbb{E}$.

A *measure* is a function: $measure : \mathbb{E} \rightarrow R$. A real value is assigned to a set of experiences. Usually, the measures are defined normalized to the $[0,1]$ interval

$$measure : \mathbb{E} \rightarrow [0,1]$$

In this paper we consider normalized mesures. This decision is not critical because each measure select the same attribute in both versions (normalized or not).

3 THE *b* MESURES

For $i=1, \ldots, k$

Let $p_i = \dfrac{\| \{ e \in E \mid C(e) = i \} \|}{\| E \|}$

the relative frequency of experiences that belong to class i in the set E.
Let $p_{1^o} , p_{2^o} , p_{3^o} , \ldots , p_{k^o}$ the p_i values in decreasing order.
We define $p_{max} = max \{ p_1 , p_2 , \ldots , p_k \}$ ($p_{max} = p_{1^o}$)

The measure $b1$ is defined as follows:

$$b1 : \mathbb{E} \rightarrow [0,1]$$
$$b1(E) = 1 - p_{max}$$

The measure $b2$ is defined as follows:

$$b2 : \mathbb{E} \rightarrow [0,1]$$
$$b2(E) = 1 - sum$$

where $sum = p_{max} - p_{max}^2 + \sum (p_i)^2$

The measure $b3$ is defined as follows:

$$b3 : \mathbb{E} \rightarrow [0,1]$$
$$b3(E) = 1 - supot$$

where $supot = \sum_{n=1}^{k} (p_{n^o})^n$

If all experiences in E belong to the same class then the values of these three measures are zero (the set E is totally ordered).

$b1$ indicates the fraction of experiences erroneously classified if the prediction of the majority class is considered.

$b2$ takes into account the prediction of the majority class and the dispersion of erroneous experiences in the remainder class. There is more order when the sizes of the classes are more unequal.

$b3$ is a weighted measure. For big classes the weight is bigger.

These measures can be used to obtain a decision tree. A criterion to select the attribute for each node is necessary. We define the *gradient* function, *Ramos-Jiménez (1999).*

Given a *measure* , *measure'*: $\mathbf{X} \times \mathbf{E} \rightarrow [0,1]$ is defined as follows:

$$measure'(X_i, E) = \sum_{j=1}^{m_i} (\| E_j \| \; measure'(E_j)) \; / \; \| E \| \quad \text{where } E_j = \{ e \in E \mid X_i(e) = j \}$$

Then, the *gradient function*, $\Delta : \mathbf{X} \times \mathbf{E} \rightarrow [0,1]$ is defined by:

$\Delta(X_i, E) = measure(E) - measure'(X_i, E)$

By considering the measures *b1* or *b2* or *b3*, the measures obtained are respectively: *b1'* or *b2'* or *b3'*.

The attribute that maximizes the gradient function (for the considered measure) is the selected attribute.

All the *b*eta measures are easy to compute, especially the first one. So these measures are interesting when the time of learning is considered critical (learning in real time) or when the data set is very big (Terabytes).

Next section shows that these measures produce good results also.

4 EXPERIMENTAL RESULTS

We have applied the TDIDT algorithm, without pruning, by using four different measures: classical Entropy (ID3) and the three new measures. The application is carried out about four standard experiences set.

These sets are *wdbc*, *car*, *pima-diabetes* y *tic-tac-toe*, that can be obtained from *MLRepository*. The following table is a brief resume of their characteristics.

Name	Cardinal	Attributes	Types	Classes	Subject
Wdbc	569	30	Numerical	2	Cancer
Car	1728	6	Symbolic	4	Automobile
pima-diabetes	768	8	Numerical	2	Diabetes
tic-tac-toe	958	9	Symbolic	2	Game

Each numerical attribute has been divided in several intervals of similar size according to the range of values.

The experimental results have been obtained by splitting each set in a random way: 80% for the training set and 20% for the test set. This process is carried out ten times for each set and for each measure. The average of the success index (SI), the number of rules and the number of nodes are shown in the table.

	SI	Rules	Nodes
Entropy	93.415	45.9	124.8
b1	91.837	65.0	170.3
b2	93.503	49.5	130.7
b3	93.503	49.5	130.7

wdbc

	SI	Rules	Nodes
Entropy	90.947	223.3	314.3
b1	82.954	367.4	518.2
b2	90.831	223.2	313.9
b3	90.717	222.5	313.4

car

	SI	Rules	Nodes
Entropy	63.565	170.5	353.9
b1	61.555	183.1	386.8
b2	63.37	171.0	357.2
b3	63.37	171.0	357.2

pima-diabetes

	SI	Rules	Nodes
Entropy	83.518	173.1	272.7
b1	80.361	206.4	328.1
b2	83.382	173.6	270.9
b3	83.382	173.6	270.9

tic-tac-toe

b2 and *b3* have an SI as good as *Entropy* in all sets.
b1 has SI values near to *Entropy* in three of four sets; the central advantage is simplicity of computation.

5 CONCLUSIONS

A new family of measures based upon the "***b***est class" concept has been defined. *b1* has a very easy computation. This easy computation can be more important than a better SI in some applications. *b2* and *b3* can be computed more easily than *Entropy*, and the sucess indexes are very similar, with a similar number of rules.

For these reasons we think that this family of measures must be borne in mind to decide the best measure for a concrete problem.

REFERENCES

Buntine, W. y Nibblett, T. A Further Comparison of Splitting Rules for Decision-Tree Induction. *Machine Learning* 8: 75-85. (1992).

Cuena, José. *Inteligencia Artificial: Sistemas Expertos*. Alianza Editorial. (1987).

Fayyad, Usama, Irani, Keki. Technical Note. On the Nandling of Continuous-Valued Attributes in Decision Tree Generation. *Machine Learning*. 8, 87-102. (1992).

Hunt, E. B., Marin, J. and Stone, P. T. *Experiments in induction*. Academic Press. (1996).

López de Mántaras R.. Technical Note. A Distance-Based Attribute Selection Measure for Decision Tree Induction. *Machine Learning*. 6, 81-92. (1991).

López de Mántaras R.. El problema de la selección de atributos en aprendizaje inductivo: Nueva propuesta y estudio experimental. *Nuevas tendencias en Inteligencia Artificial*. Universidad de Deusto. (1992).

Michalski, R. Unifying principles and a methodology of Inductive learning. *Artificial Intelligence*. (1983).

MLRepository http: //www.ics.uci.edu/˜mlearn/MLRepository

Quinlan, J.R. Discovering rules from large collections of examples: a case study. *Expert Systems in the Micho Electronic Age*. Edinburgh University Press. (1979).

Quinlan, J.R.. Induction of Decision Trees. *Machine Learning* 1: 81-106. (1986).

Quinlan, J.Ross. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc. (1993).

Ramos Jiménez, G. , Morales Bueno, R. *Formalización de los algoritmos TDIDT y CIDIM*. Informe Técnico de Investigación LCC-ITI 99/01. Departamento de Lenguajes y Ciencias de la Computación. Universidad de Málaga. (1999).

Schlimmer, J. C., Ficher, D.. A case study of incremental concept induction. *Proc. of the Fifth National Conference on Artificial Intelligence*. Morgan Kaufmann Publishers, Inc. 496-501. (1986).

Shannon, C. E. The mathematical theory of communication. *The Bell Systems Technical Journal* 27: 379-423, 623-656. (1948).

Utgoff, Paul. Incremental Induction of Decision Trees. *Machine Learning*. 4, 161-186. (1989).