Decision trees on interval valued variables

Chérif Mballo[1,2] and Edwin Diday[2]

[1]ESIEA Recherche, 38 Rue des Docteurs Calmette et Guerin
53000 Laval France
mballo@esiea-ouest.fr

[2]LISE-CEREMADE Université Paris Dauphine Place du Maréchal de Lattre
de Tassigny 75775 Paris Cedex 16. France
diday@ceremade.dauphine.fr

Submitted: June, 2004; Reviewed: February , 2005; Accepted: June, 2005

Abstract

In symbolic data analysis, the values of the variables can be, among others, intervals. The algebraic structure of these variables leads us to adapt dissimilarity ones to be able to study them. The Kolmogorov-Smirnov criterion is used as a test selection metric for decision tree induction. For this criterion, the values taken by the explanatory variables have to be ordered. We have been interested to different possible orders of these interval values. In this paper, we study, on the one hand, the reliability of the Kolmogorov-Smirnov criterion, and, on the other hand, we compare this criterion with the classical criterion of Gini.

1 Introduction

In the last few years, the extension of the classical data analysis methods to new objects with more complex structure has been an investigated topic ([2]). The aim of symbolic data analysis is to consider the particularities of the data in the treatment methods without losing information.

A classification tree is a binary tree that given an input vector of description and produces an output decision function that approximates a discrete random variable of interest called assigned function (stochastically related to the input vector of description). In our case, all the values taken by the input vector of description are interval type. Perinel ([10]) has been interested in building decision trees for symbolic variables of interval type using the criterion of Gini, the gain of information and the likelihood. The selected point for the cutting for him is the middle of the interval (a real number as in the classical case). We use the Kolmogorov-Smirnov ([7], [13]) criterion to build decision tree on interval variables ([8], [9]) so our selected point for the cutting is an interval (not the middle of the interval). The aim in building decision trees is to discriminate individuals (by their description) among k classes in several steps which correspond to the levels and nodes of the tree. The criterion of Kolmogorov-Smirnov (noted KS in the following) requires that

individuals have to be ordered by their description (the values of the explanatory variables). This criterion is used as a test selection metric (splitting criterion or attribute selection metric) to build a decision tree induction ([7]). It is based on the cumulative distribution functions of the prior classes. Its advantage is that it gives the best cutting of each explanatory variable in order that the two probability distribution functions associated to the two prior classes (or super classes obtained by regrouping of the prior classes ([3])) have the largest KS criterion value. We have been interested to different possible orders of the variables with interval values. We consider a population of n objects. Each object is described by a symbolic data table with p variables : one variable to explain (a nominal variable or class variable) and the (p-1) remaining variables are the explanatory variables (all these explanatory variables are interval type).

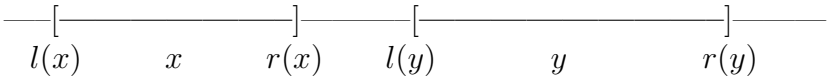
As we can order intervals in different cases, we are now in exploratory case (we explore each interval order). This method consists of building a tree for each interval order. We study the reliability of the KS criterion for each interval order. We finish this paper by the comparison of this criterion with the classical criterion of Gini.

2 Ordering interval data

In this section, we recall some methods proposed by ([4]) to order interval values. Let Ω be the set of individuals. We consider that a variable Y is called an "interval variable" or "interval valued variable" if $Y(w) = [\alpha, \beta]\ \forall w \in \Omega$, where α and β are two real numbers such that $\alpha \leq \beta$. The description of each individual (or object) is a closed interval of real numbers. We denote the set of closed intervals of real numbers by $\Im$. Thus, if $x \in \Im$, then $x = [l(x), r(x)]$ for some real numbers $l(x)$ ("l" as left bound), $r(x)$ ("r" as right bound) such that $l(x) \leq r(x)$. Taking two intervals $x = [l(x), r(x)]$ and $y = [l(y), r(y)]$, by $x = y$,we mean, of course, that $l(x) = l(y)$ and $r(x) = r(y)$. We define an interval order as an anti-reflexive and transitive binary relation defined on the set $\Im$ (notice that an interval order is a strict order, since it is not reflexive). To study an order relation on $\Im$, it is important to distinguish between two different situations: the case in which the interval are disjoint or non disjoint. Some authors ([5], [6], [11]) have been interested in interval orders. Taking two intervals x and y as indicated previously, x is considered "before" y if and only if $r(x) < l(y)$. We see that this ordering method is not "true" when the intervals are non disjoint. However, the case of non disjoint intervals have been investigated by ([12]) but in the modelling of the preferences. These authors define three binary relations P (strict preference), Q (weak preference) and I (indifference). With a finite set A, they define the necessary and sufficient conditions to associate, for each element of A, an interval in such a way that: if an interval is "completely to the right" of another interval, we obtain the relation P; if an interval is "included" into another interval, we obtain the relation I; if an interval is "to the right" of another interval but their intersection is not empty, we obtain the relation Q. We broach this problem of non disjoint intervals but in another direction that permits us to define two relations and each of them define a strict total interval order.

2.1 Disjoint intervals

Let D the subset of $\Im$ composed of disjoint intervals. If x_i and x_j are two elements of D, we have: $x_i \cap x_j = \phi \forall i \neq j$. Let x and y two elements of D. By xRy, we mean that the interval x is "*strictly before*" the interval y. We use the statement "*strictly*" to point out that the intervals are disjoints. Mathematically, this relation R on D may be defined as follows: $xRy \Longleftrightarrow r(x) < l(y)$. The relation R is transitive and anti-reflexive and define a strict total order on D. Graphically, when xRy, the interval x has a position which is "totally before" the interval y. For example, ordering two intervals x and y for a time variable, x will be "totally before" y because x" finishes before" y "starts".



2.2 Non disjoint intervals

Consider two non disjoint closed intervals of real numbers x and y. In analogy to the case of disjoint interval, we use the statement "*almost*" to point out that the intervals are non disjoints. Different cases may occur according to the position of both of the bounds of the intervals.

- Ordering "*by the lower bound*": if the lower bounds are distinct, the intervals order is related to the position of these lower bounds and if they are equal, the intervals order is related to the position of the upper bounds. By xIy, we mean that the interval x is "*almost before*" the interval y. In this case, we can remark that not every element in x is less than or equal to any element in y. For the interval x, we may say only that there is at least one element which is less than or equal to any element in the interval y. Mathematically, this relation I on $\Im$ may be defined as follows:

if $l(x) \neq l(y)$, then $xIy \Longleftrightarrow l(x) < l(y)$; if $l(x) = l(y)$, then $xIy \Longleftrightarrow r(x) < r(y)$.

The relation I is transitive and anti-reflexive and define a strict total order on $\Im$.

Considering again the case of a time variable, an interval x is "*almost before*" an interval y if x "starts before" y. In this case, it doesn't matter which one "finishes first". On the contrary, in the case in which x and y "start" at the same time ($l(x) = l(y)$), the interval which "finishes first" will be considered "*almost before*" the other one.

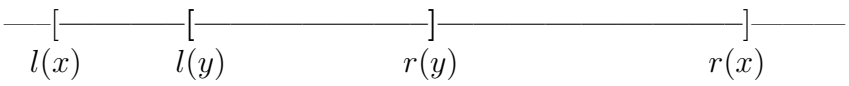
- Ordering "*by the upper bound*": if the upper bounds are distinct, the intervals order is related to the position of these upper bounds and if they are equal, the intervals order is related to the position of the lower bounds. By xSy, we mean that the interval x is "*almost after*" the interval y. In analogy to the previous situation, we can remark that not every element in y is greater than or equal to any element in x. Mathematically, this relation S on $\Im$ may be defined as follow:

if $r(x) \neq r(y)$, then $xSy \Longleftrightarrow r(x) < r(y)$; if $r(x) = r(y)$, then $xSy \Longleftrightarrow l(x) < l(y)$.

The relation S is transitive and anti-reflexive and define a strict total order on $\Im$.

In the case of a time variable, an interval y is "*almost after*" an interval x if y "ends after" x but it doesn't matter which one "starts first". On the contrary, in the case in which y and x "end" at the same time ($r(x) = r(y)$), the interval which "starts after" will be considered "*almost after*" the other one.

Remark : Let x and y two intervals such that $l(x) < l(y)$ and $r(y) < r(x)$ (the interval y is strictly included into the interval x).



Ordering by the relation I, the interval x is "*almost before*" the interval y because $l(x) < l(y)$. Ordering by the relation S, the interval x is also "*almost after*" the interval y because $r(x) > r(y)$. For any other configuration (except the case where intervals are strictly included one into another as the previous configuration), if an interval x is "*almost before*" an interval y, then the interval y is "*almost after*" the interval x (in other words, I and S create the same relation of "precedence" between two intervals). The main difference between the two orderings I and S is the case where the intervals are strictly included one into another. Taking into account this difference between orderings I and S, the reader has to decide which one is more suitable for his own data.

An alternative simplest way is to order intervals by their centers (means) or by their span lengths. The mean and the span length of each interval x is computed as follows: $center = \frac{l(x)+r(x)}{2}$ and $span\_length = r(x) - l(x)$.

3 The KS's criterion for decision tree

To build a decision tree, we need a criterion in order to separate the population of a node, into two sub populations more homogeneous (two children nodes). In this section, we study the reliability of the KS criterion according to the different interval orders. But we start by a presentation of this criterion.

3.1 Presentation of the KS's criterion

The KS splitting criterion has been introduced by Friedman ([7]) for a binary partition. Friedman supposes that the misclassification costs for the prior classes are identical, and that the prior probabilities of drawing an instance from a particular class are also identical. This criterion allows the separation of a population into two homogeneous groups. It uses the two cumulative distribution functions resulting from the grouping of the k prior classes into two classes called super classes. The method called "*towing splitting process*" ([3]) is used to generate, from the k classes, two super classes C_1 and C_2 which are associated two cumulative distribution functions F_1 and F_2 of a symbolic random explanatory variable. We can use the cumulative distribution function because we have a total order of the intervals (the elementary events are the closed intervals of real numbers). At each node, the "*towing splitting process*" method considers all the possible cases for grouping m classes ($m \leq k$) into two super classes ($2^{m-1} - 1$ possibilities). For example, if $m = 3$ to a node, we have three possibilities to group these three classes into two super classes to this node:

- $C_1 = \{\text{class 1}\}$ and $C_2 = \{\text{class 2, class 3}\}$;
- $C_1 = \{\text{class 1, class 2}\}$ and $C_2 = \{\text{class 3}\}$;
- $C_1 = \{\text{class 1, class 3}\}$ and $C_2 = \{\text{class 2}\}$.

All objects whose class is in C_i are assigned class i $(i = 1, 2)$ at each node during the splitting process. This exponential complexity $(2^{m-1}-1)$ has been reduced in a polynomial complexity by Asseraf ([1]). Let Ω the set of objects (individuals). Each object of Ω is described by p symbolic interval valued variables $X_1, X_2, ..., X_p$ (the explanatory variables) and one class variable Y (the variable to explain). Let D_{X_j} the set of values (the domain) of the explanatory variable X_j ($D_{X_j} \subseteq \Im$ where $\Im$ is the set of closed intervals of real numbers). We need to find a single cut point $x \in D_{X_j}$ that partitions the values of an interval valued explanatory variable X_j into two blocks: one block (left node) contains those values y of D_{X_j} satisfying the condition $y \leq x$ (where " $\leq$ " is an interval order) and the other block (right node) contains the remaining values as in the classical case. Let F_i^j the cumulative distribution function associated to the explanatory variable X_j for the super class C_i ($i = 1, 2$ and $j = 1, 2, ..., p$). These cumulative distribution functions are not known in practice, we only consider approximations. We can consider the approximate cumulative function because we have an order and the set $\{y \in D_{X_j}/y \leq x\} \cap \{y \in D_{X_j}/y \in C_i\}$ is always finite in practice. Let $\hat{F_i^j}$ the approximate cumulative function of F_i^j. For each explanatory variable X_j and for each interval value $x \in D_{X_j}$, the approximation $\hat{F_i^j}$ which estimates F_i^j in $x \in D_{X_j}$ is defined by:

$$\hat{F}_i^j(x) = \frac{cardinal(\{y \in D_{X_j}/y \leq x\} \cap \{y \in D_{X_j}/y \in C_i\})}{cardinal(\{y \in D_{X_j}/y \in C_i\})} \quad (1)$$

So the KS's criterion (or an optimal cut point x) is defined by:

$$KS = \sup_{x \in D_{X_j}} \left| \hat{F_1^j}(x) - \hat{F_2^j}(x) \right| \forall j = 1, 2, ..., p \quad (2)$$

We see by these two previous formulas that the selected argument for the cutting is a closed interval and not a real number as in the classical case. Thus, we can use all the other steps (because these steps are common to all the variables) in order to build the decision tree. The principle consists of selecting, at each node, the most discriminating explanatory variable X_j (therefore the corresponding interval x for the cutting).

3.2 Comparison

As we have different possibilities to order intervals, we must examine the results obtained for each order to study the reliability of the KS criterion. We have been interested to the misclassification rate of the test set. To estimate this parameter, we use the hold out method ($\frac{1}{3}$ of the objects for the test set and $\frac{2}{3}$ for the learning set for each database used). The aim of this section is to explore the different interval orders. We compare, on the one hand, these orders using the KS splitting criterion, and on the other hand, the two criteria (KS and Gini) for each interval order. The table (Table 1) present the databases used (these databases are available at *www.ceremade.dauphine.fr/ touati/exemples.htm*). From left to right of this table, the columns indicate respectively: the number of the files, their names, their size (number of individuals or objects), the number of prior classes of the variable to explain (the class variable), the distribution of the prior classes and the number of the explanatory variables (the predictors). For the column "*Distribution-classes*", for

example the notation (5;3) indicates that five objects are in class "1" and three objects in class "2" for a class variable having two classes. To stop the splitting process, we have some parameters: "minimal size of a terminal node" (this parameter varies according to the size of the file) and "minimal size to split a non pure node" (this parameter is twice the first). We use the pure affectation. We present the results for all interval orders (lower bound, upper bound, mean and span length) because we are now an exploratory situation (explore all the interval orders).

Number	*Name*	*Size*	*Nb-cl*	*Distribution-classes*	*Nb-Pred*
F1	Travel	21	5	(5;4;3;6;3)	2
F2	Wine	23	9	(2;2;2;2;6;5;2;1;1)	21
F3	Player	29	2	(11;18)	7
F4	Iris-Fisher	30	3	(10;10;10)	4
F5	Wave	30	3	(10;10;10)	21
F6	Auto	33	4	(10;8;8;7)	8
F7	Football	45	4	(16;21;7;1)	7
F8	Accident	48	3	(36;9;3)	5
F9	Temperature-1988	60	6	(14;12;11;4;10;9)	12
F10	Shuttle	102	7	(78;1;1;15;5;1;1)	9
F11	Cholesterol	193	2	(103;90)	2
F12	Age-color	231	2	(108;123)	10
F13	Glucose	690	2	(335;355)	2
F14	Prof-size	720	5	(36;441;32;35;176)	1
F15	Temperature-74-88	900	2	(443;457)	12
F16	Regions-UN	10000	4	(1900;2300;3620;2180)	4

Table 1. Inventory of the databases used

The figures (Figure 1 and Figure 2) presents the results obtained by the criteria of KS and Gini for each interval order. For each figure, the notations "Upper, Lower, Mean, Span Length" indicate respectively that the intervals have been ordered by the "upper bound, lower bound, mean, span length". Despite some similarities (F9, F16 for figure 1 and F10, F15, F16 for figure 2), these results show that the rate of misclassification often differs strongly from one file to another and from one interval order to another. With these results, we see that there is not an interval order that is better than others.

To compare the two criteria KS and Gini, we must examine the results of these two figures (figure 1 and figure 2) but by order. The figures (figure 3 to figure 6) present these results. The table (Table 2) summarizes these results. We have many files where the two criteria KS and Gini present the same results (column "*Equality*"), especially for the order by the lower bound (figure 4). Generally, we remark that these results are very dispersed according to the files and the interval orders. The last file (F16) presents a marked advantage in comparison with the others files (its misclassification rate is very small).

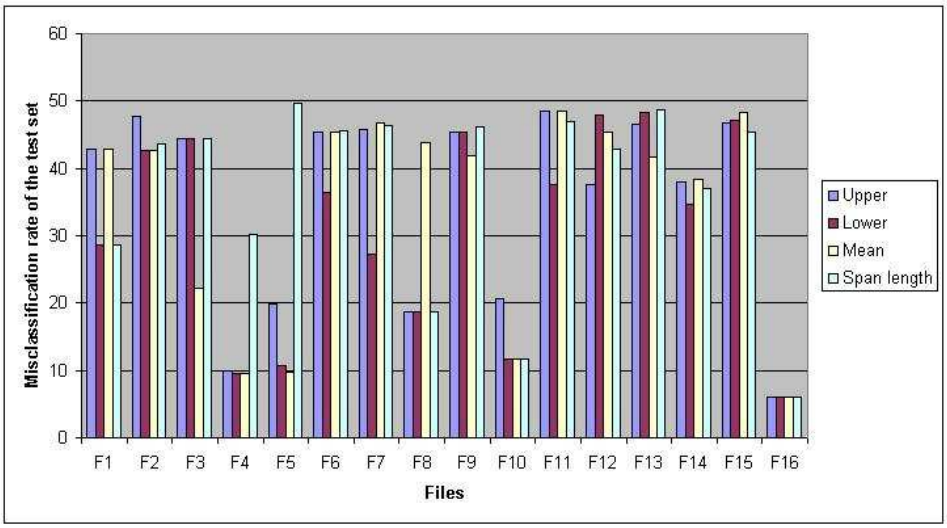


Figure 1: Results obtained by the KS's criterion

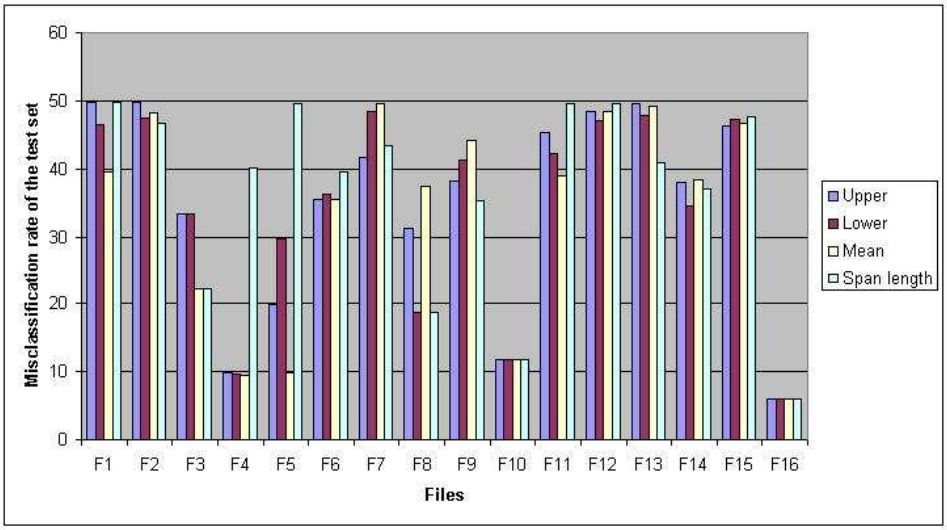


Figure 2: Results obtained by the criterion of Gini

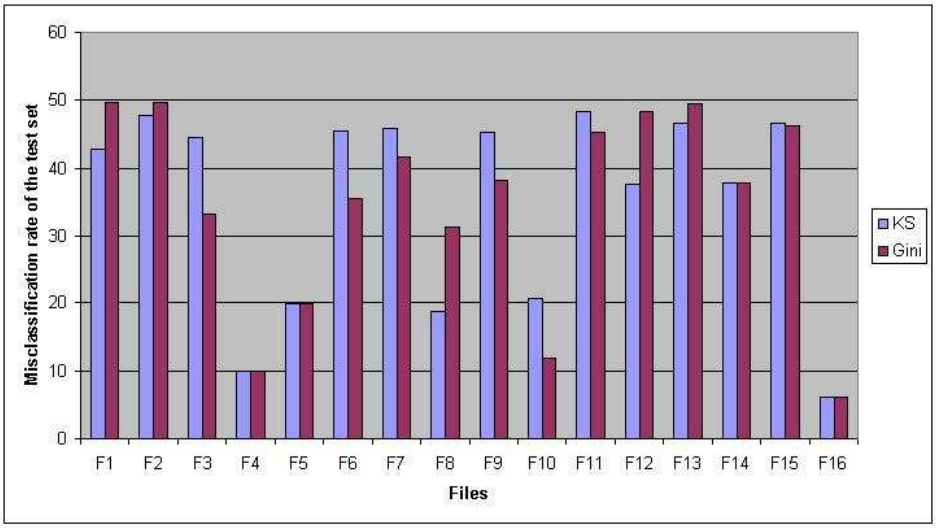


Figure 3: Results obtained by the order by the upper bound

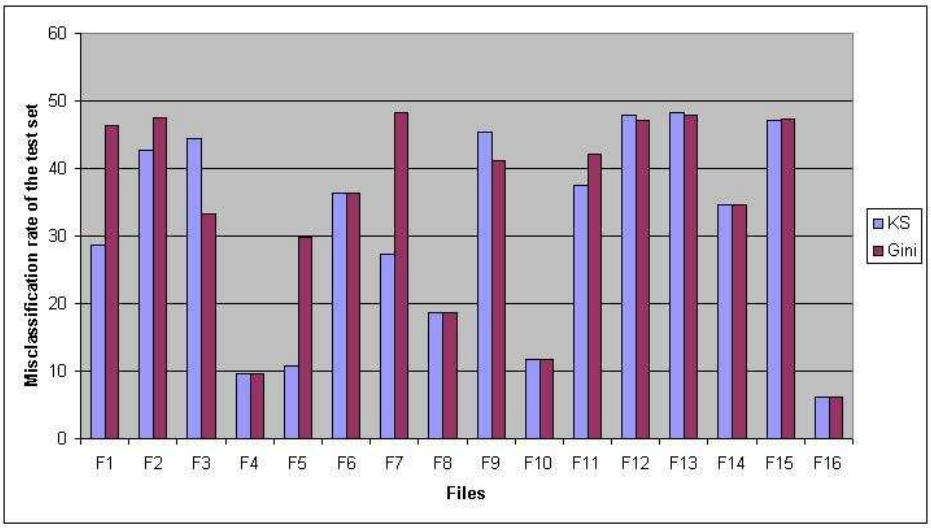


Figure 4: Results obtained by the order by the lower bound

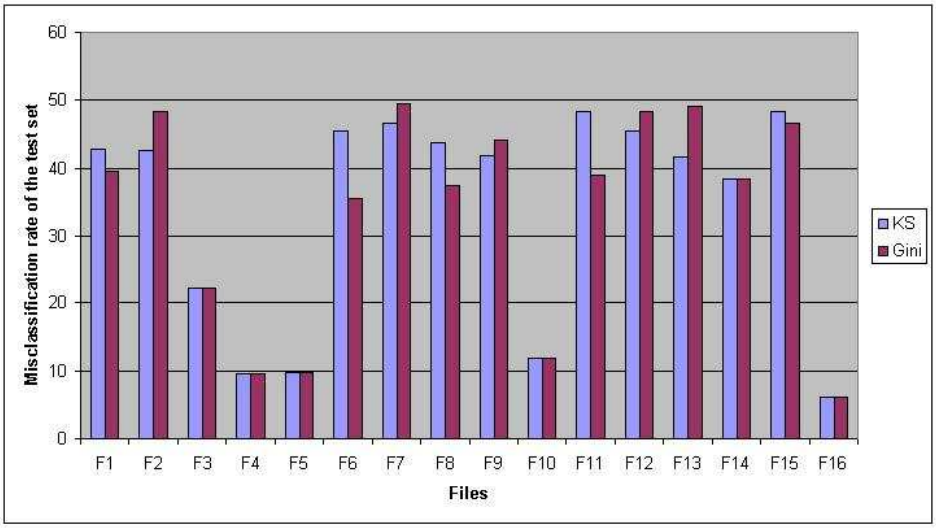


Figure 5: Results obtained by the order by the mean

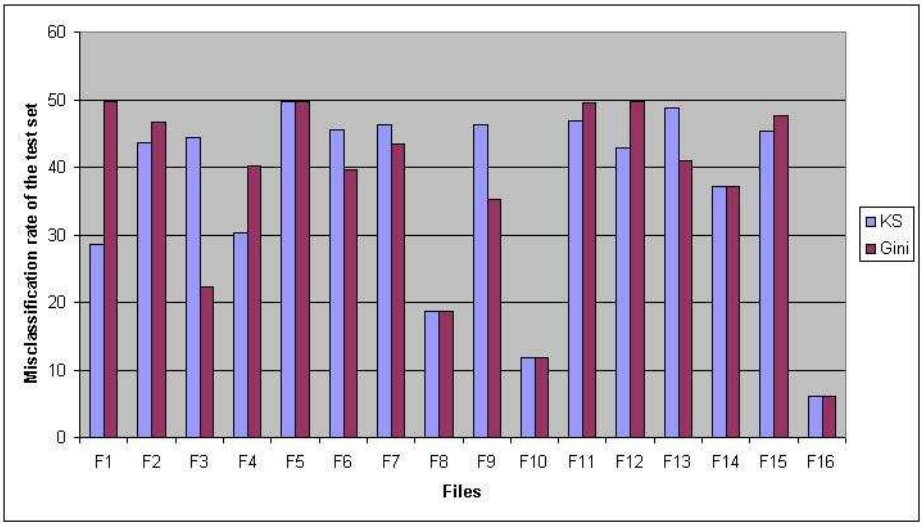


Figure 6: Results obtained by the order by the span length

	Files where KS is better	*Files where Gini is better*	*Equality*
Figure 3	F1,F2,F8,F12,F13	F3,F6,F7,F9,F10,F11	F4,F5,F14,F15,F16
Figure 4	F1,F2,F5,F7,F11	F3,F9	F4,F6,F8,F10,F12, F13,F14,F15,F16
Figure 5	F2,F7,F9,F12,F13	F1,F6,F8,F11,F15	F3,F4,F5,F10,F14,F16
Figure 6	F1,F2,F4,F11,F12,F15	F3,F6,F7,F9,F13	F5,F8,F10,F14,F16

Table 2. Recapitulative results of the figures (figure 3 to figure 6)

4 Conclusion and perspectives

In this paper, we point out the extension of the KS's criterion to the case where the explanatory variables are interval type. This criterion treat all the explanatory variables together during the splitting process and it deduces the more discriminating for each interior node. Our method explore all the interval orders because we have different possibilities to order intervals. We compared this criterion with the classical criterion of Gini and the results obtained are very dispersed. We propose in our future work to study the stability and the robustness of the KS's criterion on interval data.

The aim in building decision tree by the KS's criterion is to extract symbolic objects from the decision tree and to induce from these symbolic objects a new data table in order to study them by a symbolic data analysis of higher level.

References

[1] ASSERAF, M. (1998): Extension et optimisation pour la segmentation de la distance de Kolmogorov-Smirnov. Ph.D Thesis, Université Paris Dauphine, Paris.

[2] BOCK, H. H., DIDAY, E. (2000): *Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data.* Springer, Berlin-Heidelberg.

[3] BREIMAN, L., FREIDMAN, J. H., OHLSEN, R. A., STONE, C. J. (1984): *Classification and Regression Trees.* Belmont, CA Wadsworth.

[4] DIDAY, E., GIOIA, F., MBALLO, C. (2003): Codage qualitatif d'une variable intervalle. *Journées de Statistique*, **35**, 415-418.

[5] EIJGENRAAM, P. (1981): *The solution of initial value problems using interval arithmetic formulation and analysis of an algorithm.* Mathematisch Centrum, Amsterdam.

[6] FISHBURN, P. C. (1985): *Interval orders and interval graphs: A study of partially ordered sets.* A Wiley-Interscience Publication.

[7] FRIEDMAN, J. H. (1977): A recursive partitioning decision rule for non parametric classification. *IEEE Transactions on Computers*, **C-26**, 404-408.

[8] MBALLO, C., ASSERAF, M., DIDAY, E. (2004): Binary decision trees for interval and taxonomical variables. *A Statistical Journal for Graduates Students (incorporating Data and Statistics)*, **Volume 5, Number 1**, April 2004, 13-28

[9] MBALLO, C., DIDAY, E. (2004): Kolmogorov-Smirnov for decision tree on interval and histogram variables. *In Studies in classification, Data Analysis and Knowledge organization: Classification, Clustering and Data Mining Applications*, editors: D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul; Springer; Proceedings of the *International Federation of the Classification Societies*, Chicago, USA, July, 15-18, 2004, 341-350.

[10] PERINEL, E. (1996): Segmentation et Analyse des données symboliques : Application á des données probabilistes imprécises. Ph.D Thesis, Université Paris Dauphine, Paris.

[11] PIRLOT, M., VINCKE, PH. (1997): *Semi-orders: properties, representations, applications.* Kluwer Academic Publisher.

[12] TSOUKIAS, A., THE, N. A. (2001): Numerical representation of PQI interval orders. *LAMSADE Université Paris Dauphine*, **184**, 1-27.

[13] UTGOFF, P. E., CLOUSE, J. A. (1996): A Kolmogorov-Smirnoff metric for decision tree induction. *University of Massachusetts Amherst*, **96-3**, 1-10.