Dynamic Classification Trees for imprecise data

Massimo Aria and Valentina Cozza

Abstract This paper provides a supervised classification tree-based methodology to deal with Multivalued data, specifically predictors measurements can be provided by a functional distribution or an interval of values. Main literature refers to symbolic data analysis, aiming to extend standard methods such as factorial analysis, clustering, discriminant analysis, etc., to deal with symbolic data tables. One approach is to define a suitable data pre-processing enabling the application of standard methods. A more correct approach is to define suitable methods to deal specifically with unstandard data. In the framework of supervised classification, there are no proposal in literature for supervised classification methods to deal with both standard and multivalued data as well. There are only proposals based on data pre-processing. This paper provides a methodology to grow the so-called Dynamic CLASSification TREE (D-CLASSTREE), upon suitable definition of both a specific splitting criterion and a tree-growing algorithm. A real world case study will be considered to show the advantages of the final output and main issues of the interpretation. A comparative study with older proposals will be also described such to demonstrate the stability and the better accuracy of the D-CLASSTREE.

Key words: bla

1 The imprecise data

The results of measurements are not precise numbers or vectors but more or less imprecise numbers or vectors (Viertl, 1999, 2003). Imprecision is different from measurement errors or stochastic uncertainty. In the last decade, the literature about the statistic treatment of imprecise measurement counts several contribution (Couso and Sanchez, 2011; Ferraro, Colubi, Gonzalez-Rodriguez, and Coppi, 2011; Gil et

Department of Mathematics and Statistics, University of Naples Federico II, e-mail: (massimo.aria, valentina.cozza)@unina.it

al., 2006). Special cases of imprecise data are both interval and histogram data. They are typical where training data comes with intrinsic uncertainty that might be the result of imprecise measuring instruments, as in image recognition framework, human judgments, etc. We intend our approach as a 'subjectivist' view of imprecision formalizing the uncertainty concerning an underlying 'crisp' phenomenon.

2 Tree-based methods

Data can be hierarchically organized in a connected and oriented graph, the so-called tree, characterized by a set of linked nodes, in which any two nodes are connected by exactly one simple path, the starting-node is the *root* and the end-nodes are the *leaves*. Two properties are satisfied: the *shape property*, where each node has a fixed number r of child nodes (for $r = 2$ it is assumed a binary tree); the *heap property*, where each node is greater than or equal to each of its children according to some comparison predicate which is fixed for the entire data structure. Trees have been used in supervised classification and non parametric regression. Each node of the tree includes statistical units or objects which are recursively partitioned such to reduce the impurity of a target or response variable as explained by a set of available predictors. To each leave of the tree is assigned a response value/class, the set of leaves describes a partition of the given sample of objects, each path of the tree gives the sequential conditions of the predictors measurement which is necessary to belong to each final leave. In such a tree graph, a new object for that only the predictors measurements are known can be slide down until one of the leaves where it is possible to predict its response value/class on the basis of the prior leave's assignment done in the tree growing. The quality of the prediction can be evaluated in terms of misclassification rate or mean square error estimates based on learning sample (too optimistic), test sample (which requires large sample size), cross-validation (for small sample size).
Main focus of recent literature is to outperform the decision/prediction rule in terms of accuracy such to answer the bias-variance dilemma with alternative solutions. Enhancements are provided by ensemble methods, random forest, evolutionary programming. All these approaches do not provide one tree structure for prediction denying the interpretability advantage of the tree graph to describe the hierarchical dependence relationships. The final assignment of one object is induced by a suitable combination of tree structures. Ensemble methods are learning algorithms that develop a population of simple models (like trees), called weak learner, from the perturbed training set combining them to form a composite predictor, which is generally more accurate than the single trees whence it is formed by. Ensemble of classifiers works by constructing a set of weak learners and then classifying new data points by taking a vote of their predictions. Even though there exist several ways to build ensemble (Dietterich, 2000), the most popular ensemble methods, such as Bagging (Breiman, 1996), Boosting (Freund and Schapire, 1997) and Random Forest (Breiman, 2001), work by manipulating the training examples through

re-sampling methods. All of these algorithms aggregate the object decisions by voting, but none of these ensemble methods allows to preserve the final tree-structure: if we are interested in the accuracy of the prediction then we can use an ensemble because it is absolutely more accurate than a single decision tree, but the interpretation of the the tree-structure is irreparably lost because the aggregation process compromises the construction of a unique prediction tree structure.

3 The proposed methodology for multiple values data

3.1 The multiple values data description

Multiple values variables (MVV) are included in the category of symbolic data (Bock and Diday, 2000). The data descriptions of the units are called *symbolic* when they are more complex than the standard ones due to the fact that they contain internal variation and are structured. Symbolic data need more complex data tables called *symbolic data tables* because a cell of such data table does not necessarily contain as usual, a single quantitative or categorical values. The symbolic variables are usually represented as weight (probability) distributions or interval values. Let $\mathbf{X}$ be a continuous variable defined on a finite support $S = [\underline{x}, \overline{x}]$, where $\underline{x}$ and $\overline{x}$ are the minimum and maximum values of the domain of $\mathbf{X}$.
A histogram of $\mathbf{X}$ is the representation of an empirical distribution, described by a set of pairs (I_h, π_h), $h = 1, \ldots, H$, where H is the number of contiguous intervals (bins) $\{I_1, \ldots, I_h, \ldots, I_H\}$, where $I_h = [\underline{x}; \overline{x}[$, in which the support S is partitioned and π_h is the frequency associated with each interval.
A generic interval variable $\mathbf{X}$ is a correspondence between a set E of units and a set of closed intervals $[X_{min},\ X_{max}]$, with $X_{min} \leq X_{max}$ and $X_{min}, X_{max} \in \Re$.
In dealing with imprecise data, in the literature tree-based methods are used with interval data as predictors by Mballo and Diday (2005), and by Limam, Diday and Winsberg, (2003). A preliminary pre-processing of interval data is mandatory to build the tree-based structure. This pre-processing consists either in considering the lower bound of each interval or the upper bound of each interval. Then a normal tree-growing procedure is done by taking as impurity measure the Kolmogorov-Smirnov measure. As alternative pre-processing of interval data, the mean value of each interval can be considered. Authors does not consider the possibility to have histogram data.

3.2 The new definition of split

Tree-growing depends on the nature of both the response variable and the predictors. Response variable governs the choice of the impurity criterion as well as predictors

govern the way the splitting variables are defined. Traditionally the number of possible split to be generated by each predictor depends on the nature of the predictor itself (i.e. numerical, ordinal or nominal). As the partitioning procedure produces binary splits, in the case of numerical (ordinal) predictors with N (M) distinct values $N-1$ ($M-1$) possible binary splits can be generated. In the case of nominal predictors with M distinct modalities, $2^{M-1}-1$ possible binary splits can be generated. In our case, the predictor matrix involves both classical variables and Multiple Values Variables, so it is necessary determine how splitting variables can be generated by the latter variables.
Let Γ be a set of $n \times Q$ Multiple Valued Variables, let $\mathbf{Z}$ be a $n \times K$ random variable and let Y be a n-dimensional vector representing the response variable. Define $\mathbf{X} = [\Gamma\ \mathbf{Z}]$ as the predictors matrix of dimension $n \times P$, with $P = K + Q$. Let X^p be the p^{th} predictors, with $p = 1, \ldots, P$.
Suppose that X^p is represented by multiple values variable of the type *histogram data*.
Let $F_{X^p\_i}(u)$ be the empirical cumulative distribution function (ECDF) of the i^{th} instance of the predictor X^p, and let $F_{X^p\_j}(u)$ be the ECDF of the j^{th} instance of the same predictor. The following cases can be verified:

- $F_{X^p\_i}(u) = F_{X^p\_j}(u)$;
- $F_{X^p\_i}(u) < F_{X^p\_j}(u)$;
- $F_{X^p\_i}(u) > F_{X^p\_j}(u)$.

With respect to the distribution of the i^{th} instance, to generate splitting variables we order the instances characterized by histogram data by verifying the former inequalities via the well-known T statistics of Wilcoxon test.
The splitting ternary partition is given by a joint lecture of both T-statistics and the connected $p-value$. Consider we are using the i^{th} instance as reference instance, and we are deciding in which child node will fall down the j^{th} instance. Indeed if $T < 0$ and $p-value < \alpha$, then we are considering the first case and j^{th} instance goes down in the left child node. On the other hand, if $T > 0$ and $p-value < \alpha$, then we are considering the second case and j^{th} instance goes down in the right child node. If $p-value > \alpha$ we are considering the third case, and j^{th} instance goes down in the central child node. We can conclude that, if there are N distinct histograms the number of possible splits to be generated is equal to $N-2$.
Suppose now X^p is the p^{th} predictor in the data matrix, and it is represented by multiple values variable of the type *interval data*. Let X^{pi}_{min} and X^{pi}_{max} be respectively the lower and the upper bound of the interval of the i^{th} instance of the predictor X^p. Let X^{pj}_{min} and X^{pj}_{max} be respectively the lower and the upper bound of the interval of the j^{th} instance of the predictor X^p. With respect to the i^{th} instance, the following cases can occur:

1. $X^{pj}_{min} < X^{pi}_{min}$ and $X^{pj}_{max} < X^{pi}_{max}$;
2. $X^{pj}_{min} > X^{pi}_{min}$ and $X^{pj}_{max} > X^{pi}_{max}$;
3. $\left\{X^{pj}_{min} \geq X^{pi}_{min} \text{ and } X^{pj}_{min} \leq X^{pi}_{min}\right\}$ or $\left\{X^{pj}_{min} \leq X^{pi}_{min} \text{ and } X^{pj}_{min} \geq X^{pi}_{min}\right\}$

In the first case the j^{th} instance goes down in the left child node, in the second case j^{th} instance goes down in the right child node, in the third case j^{th} instance goes down in the central child node. As in the case of histogram data, we can conclude that, if there are N distinct intervals the number of possible splits to be generated is equal to $N-2$.

3.3 Dynamic classification trees for imprecise data

Table 1 shows the pseudo-code of the Dynamic Classification Tree for Imprecise Data.

Let Γ be a set of $n \times Q$ Multiple Valued Variables, let Z be a $n \times P$ random variable and let Y be a n-dimensional vector representing the response variable. Define $X = [\Gamma\ Z]$ as the predictors matrix of dimension $n \times P$, with $P = K + Q$. Let X^p be the p^{th} predictors, with $p = 1,\ldots,P$.
Initialize → generate root node

- While a stopping rule is not verified
  - if the current node is not terminal
    - for p=1:P
      - if is ordinal X^p || is numerical X^p,compute the decrease in impurity for all the $N-1$ splitting variables and store its maximum
      - elseif is categorical X^p, compute the decrease in impurity for all the $2^{M-1}-1$ splitting variables and store its maximum
      - elseif is histogram X^p, compute the decrease in impurity for all the $N-2$ splitting variables and store its maximum
      - elseif is interval X^p, compute the decrease in impurity for all the $N-2$ splitting variables and store its maximum
      - end
    - end
  - Generate three or two children nodes according to the nature of the predictor generating the higher decrease in impurity computed in the previous loop. Update the status of the generated child node (Internal-Terminal) and assign a number to each of them.
    - if is terminal the generated node
      - store the node
    - else
      - continue: generated child node becomes now a father node.
    - end
  - end
- end

Output → Ternary classification tree

Table 1 Pseudocode of Dynamic classification trees for imprecise data

The innovative contribution of our algorithm refers to tree-growing procedure, specifically it refers to a new way to define the splitting variables. With respect to explorative purposes, it means that the interpretability of partitions takes in account a more rigorous information when MVV predictors generate splits. The intrinsic uncertainty in pre-process such variables disappears because no pre-processing is done to perform the analysis. About decisional purposes, none is changed with respect to classical approaches. Indeed both division of the total sample in learning sample and test sample and cross-validation procedures are possible. Our approach allows to such a classifier to preserve the conditions to be used with ensemble methods such as Bagging, Boosting, Random Forests, etc. (Breiman, 1996, 2001; Freund and Schapire, 1997).

4 A real world case

The methodology have been performed on a database of the Department of Dermatology of the Second University of Naples. The database consists of 220 skin lesion dermoscopic images, for which a histological diagnosis is available, with a resolution of 768×512 pixels, divided into two classes: 86 images are relative to malignant melanoma and 134 of these lesions are classified as benign lesions. The skin lesion dermoscopic images are acquired using a charge-coupled devise camera connected to an epiluminescence microscopy.
The dataset consists in 34 variables or *descriptors* (including 11 point values, 6 intervals data and 17 histograms data), plus a binary response variable. The multi-valued data describing the dermoscopic image database is structured as a matrix $D = \{d_{i,p}\}$, where the rows represent the statistical units, i.e. the images, and the columns represent the multi-valued descriptors. Each matrix cell $d_{i,p}$ indicates the set of values attained by the i^{th} image for the p^{th} descriptor, that can be a scalar real value, an interval value, or a set of histogram values.
Following the ABCD-rule of dermoscopy (Stolz et al., 1994), descriptors chosen for characterizing different lesion classes consist of quantitative measures of asymmetry, border, and color information extracted by dermoscopic images. More details about ABCD-rule can be found, in example, in Bono et al., 1999, Celebi et al., 2007; Maglogiannis and Kosmopoulos, 2006.

4.1 The results

Figure 1 shows the tree-structure of the Dynamic Classification Tree.

The figure emphasizes the way the splits are generated. For ternary splits, if the splitting variable is generated by a histogram variable, a plot showing the Kernel density function estimate of the typical distributions is put in the graph. The central density function (in grey) refers to the distributions going down in the central child node. The left and the right density functions (respectively bold-black and dot-black) refer to the distributions going down respectively in the left and right children nodes. If the splitting variable is generated by interval data, a plot showing these intervals is put in the graph. The central interval refers to images going down in the central child node, as well as upper-left and lower-right intervals refer to images going down respectively to the left and right children node. If the splitting variable is generated by point variables, the the split is binary and in the figure is indicated the cutting point. The error rate at root node is equal to 0.3909 as well as error rate of the tree is equal to 0.1909.
Table 2 shows the DCTree in table format. First four columns indicate respectively the node number, the node size, the children nodes generated by the actual node and the father of the actual node. Column named splitting predictor indicates which predictor generates the split. In parenthesis the nature of the predictor is indicated (H if histogram, I if interval, P if point).The column named cutting point describes the

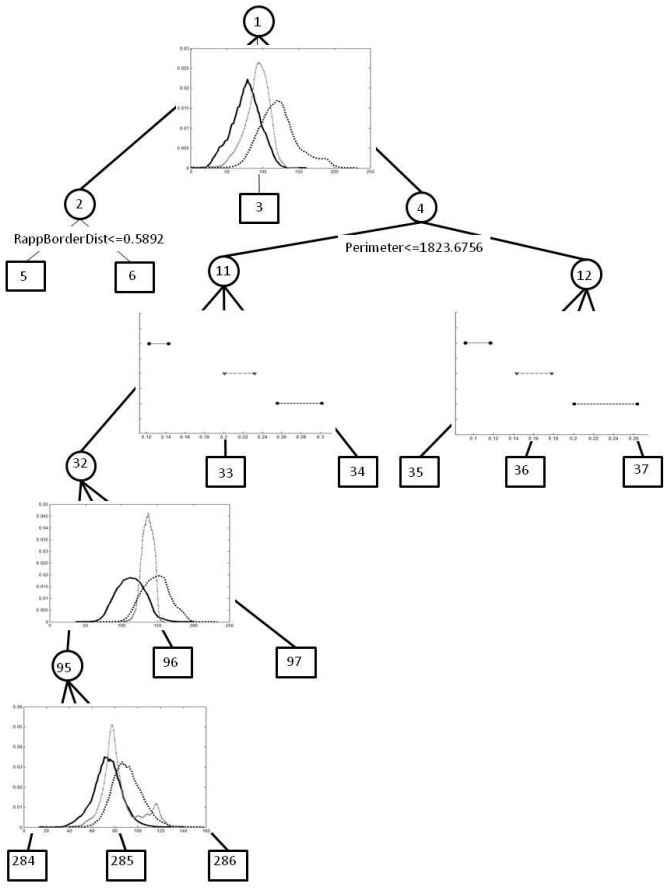

Fig. 1 Dynamic Classification tree of dermoscopic images database

split. If the splitting predictor is a histogram data, then some descriptive information about the distribution of the reference image is reported in the column (precisely Min, Max, Mean, Standard Deviation, Skewness, Kurosys). If the splitting predictor is a interval data, then upper and lower bounds of the reference interval are respectively reported in brackets in the column. If the splitting predictor is a point variable, then the cutting point is reported in the column.

The last two columns refer to the misclassification ratio within node (Rt) and to the assigned class within node.

References

1. Bock H., Diday E. (2000). *Analysis of Symbolic Data*, Springer-Verlag, Heidelberg.

Node number	Size	Children	Father	Splitting Predictor	Cutting point	Rt	Class
1	220	2 3 4	-	MagloZonaEsterna (H)	0.33 145.66 91.34 17.74 −1.07 5.51	0.39	Benignant
2	65	5 6	1	RapportoBorderDist (P)	0.589	0.20	Malignant
3	3	-	1	Terminal	-	0.00	Malignant
4	152	11 12	1	Perimeter (P)	1823.67	0.21	Benignant
5	46	-	2	Terminal	-	0.06	Malignant
6	19	-	2	Terminal	-	0.37	Benignant
11	95	32 33 34	4	AsimmXYRosso (I)	[0.20 0.23]	0.06	Benignant
12	57	35 36 37	4	AsimmXY (I)	[0.14 0.18]	0.44	Benignant
32	86	95 96 97	11	MagloZonaEsterna (H)	98.67 157.33 136.04 8.09 −0.44 3.21	0.03	Benignant
33	6	-	11	Terminal	-	0.33	Malignant
34	3	-	11	Terminal	-	0.00	Benignant
35	23	-	12	Terminal	-	0.35	Benignant
36	11	-	12	Terminal	-	0.09	Benignant
37	23	-	12	Terminal	-	0.29	Malignant
95	73	284 285 286	32	MagloZonaIntermedia (H)	38.00 133.33 83.47 15.90 0.87 3.36	0.03	Benignant
96	1	-	32	Terminal	-	0.00	Malignant
97	12	-	32	Terminal	-	0.00	Benignant
284	29	-	95	Terminal	-	0.03	Benignant
285	1	-	95	Terminal	-	0.00	Malignant
286	43	-	95	Terminal	-	0.00	Benignant

Table 2 DCTree description

2. Breiman, L. (1996). Bagging Predictors, *Machine Learning*, 26, 46-59.
3. Breiman, L.(2001). Random Forests, *Machine learning*, 45(1), 5-32.
4. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984): *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.
5. Celebi, M.E., Kingravi, H.A., Uddin, B., Iyatomi, H., Aslandogan, Y.A., Stoecker W.V., Moss R.H. (2007). A Methodological Approach to the Classification of Dermoscopy Images. *Computerized Medical Imaging and Graphics*, 31(6), 362-373.
6. Couso, I., Sánchez, L. (2011). Mark-recapture techniques in statistical tests for imprecise data. *International Journal of Approximate Reasoning*, 52, 240260
7. Dietterich, T.G. Ensemble methods in machine learning. In J.Kittler and F.Roli, editors, multiple classifier system. *First International Workshop, MCS 2000,Cagliari, volume 1857 of lecture notes in computer science*. Springer-Verlag. (2000)
8. Ferraro, M.B., Colubi A., González-Rodríguez, G., Coppi, R. (2011). A determination coefficient for a linear regression model with imprecise response. *Environmetrics*, 22, 516-529.
9. Freund, Y., and Schapire, R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1).
10. Gil, M.A., Montenegro, M., González-Rodríguez, G., Colubi, A., Casals, M.R. (2006). Bootstrap approach to the multi-sample test of meanswith imprecise data.*Computational statistics and data analysis*, 51, 148-162
11. Hastie T., Tibshirani R., Friedman J. (2008): *The Elements of Statistical Learning*, 536, Springer-Verlag.
12. Limam, M.M., Diday, E., Winsberg, S. (2003). Symbolic class description whit interval data. *Journal of Symbolic Data Analysis*, 1(1).
13. Maglogiannis, I., Kosmopoulos, I. (2006). Computatioal vision systems for the detection of malignant melaoma. *Oncology Reports*, 15, 1027-1032.
14. Mballo, C., Diday, E. (2005). Decision trees on interval valued variables. *The Electronic Journal of Symbolic Data Analysis*, 3(1), 8-18
15. Stolz, W., Riemann, A., Cognetta, A.B., Pillet, I., et al (1994). ABCD rule of dermoscopy: a new pratical method for early recognition of malignant melanoma. *European Journal of Dermatology*, 4, 521-527
16. Viertl, R. (2003). Statistical inference with imprecise data. In *Encyclopedia of Life Support Systems* (EOLSS), Eolss Publ., Oxford, www.eolss.unesco.org.