BoolFuncOnDB k-value README

Introduction

The BoolFuncOnDB program takes a k-valued or Boolean function and dataset as input, then assigns a value to each to each datapoint based on the function. The value is Boolean, so 1 or 0. If a particular attribute of the dataset is not Boolean, then the program will automatically detect this and ask the user if they would like to use ordinal k-values and define a series of thresholds (intervals) for the chosen k-value. This is needed because a function cannot be applied to the datapoints if the datapoints' values are numeric and open-ended. If the user chooses to use k-values, then the k-value must be ordinal (lesser values are confirmed to be lesser than greater values) for the process of applying a function to work. If the user chooses to not use k-values, then the program will ask the user to define a range of two thresholds, or a minimum or maximum threshold, respectively. Values of "yes," "true," "t," "1," and others are accepted as Boolean. Otherwise, that particular attribute will need a threshold or series of thresholds. In the case of some value of something like "random_text," the program will not work.

Definitions

Range: this is defined as an upper or lower bound. Therefore, a threshold of [2, 3] will define any value between 2 and 3 as true (1) and any number outside that threshold will be false.

Maximum threshold: for example, a maximum threshold of 2 will define any value less than or equal to 2 as true, and greater than 2 will be false.

Minimum threshold: for example, a minimum threshold of 2 will define any value greater than or equal to 2 as true, and lesser than 2 will be false.

K-valued thresholds: for k-valued thresholds, the thresholds are a series of intervals. For example, for the value $k = 3$,

[0, 5) = 0; [5, 10) = 1; [10, infinity) = 2.

Numeric values in the first interval are given a k-value of 0, values in the second interval are given a k-value of 1, and values in the last interval are given a k-value of 2. For a datapoint "100, 0, 5" with the assigned k-values of [3, 3, 3], the k-value representation of that datapoint will be "2, 0, 1.

Funtion value: the value for a particular datapoint that is determined by the function on that datapoint. This could also be considered assigning a class to the datapoint.

Input

The program first and foremost needs a csv file dataset as input. It can be named anything as long as the name is present in the “filename” variable, which is located at the top of the header file. The file needs to be present in the executable’s directory or the exact path needs to be provided in the filename. The first line of the file must be the names of the attributes. If there are more or less attributes than what is in the datapoints, or if the dataset is sparse, the program will not work. The dataset must be numeric, Boolean, or k-value. Numeric attributes must be assigned thresholds for the program to work. The program will ask the user if they would like to use k-value thresholds even if numeric values are not present because this must be determined beforehand. If k-value is not chosen, then Boolean thresholds will be used. K-values can also be automatically detected by the program provided that the attributes have “_kv” appended to the end of their name. For example, if a particular column “x1_kv” has a maximum value of 3, then the k-value assigned to that attribute will be 4.

Moreover, the user input for a Boolean function must be of the form in the disjunctive normal form. For example, “x1x2x3 v x4x5 v x10” will work. The format is the similar for k-value functions, but it is recommended for “AND” and “OR” statements to append the “&” or “v” symbols between the attributes. Moreover, each attribute in the function needs the k-value or k-values which are necessary in that particular function for a particular datapoint to be true. For example, “x1=0 & x2<4 v x3>1” is a valid function. The first clause states that the datapoint must have a k-value representation where the attribute x1 is equal to 1 and the attribute x2 is less than 4 for the datapoint to be assigned a value of 1. The “OR” clause simply specifies that the third attribute must be greater than 1 for the other condition for the datapoint to be assigned a value of 1. Lesser than or equal to and greater than or equal to symbols are acceptable as well, so “>=” and “<=” will work. Using parenthesis in the equation to make it more readable or in place of the “AND” symbol is acceptable. In fact, no “AND” symbol or parenthesis are necessary at all, but otherwise k-value functions can be quite confusing to look at. For example, “x1=0x2<4x3=1 v x2>=2x4>3” is not readable at all, but it will still be properly interpreted by the program. Even though functions can be entered like this, they will be printed with the proper symbols in the results file. Furthermore, if function is entered without specifying what k-values are necessary, then it will be assumed that values of 1 and greater are valid.

Output

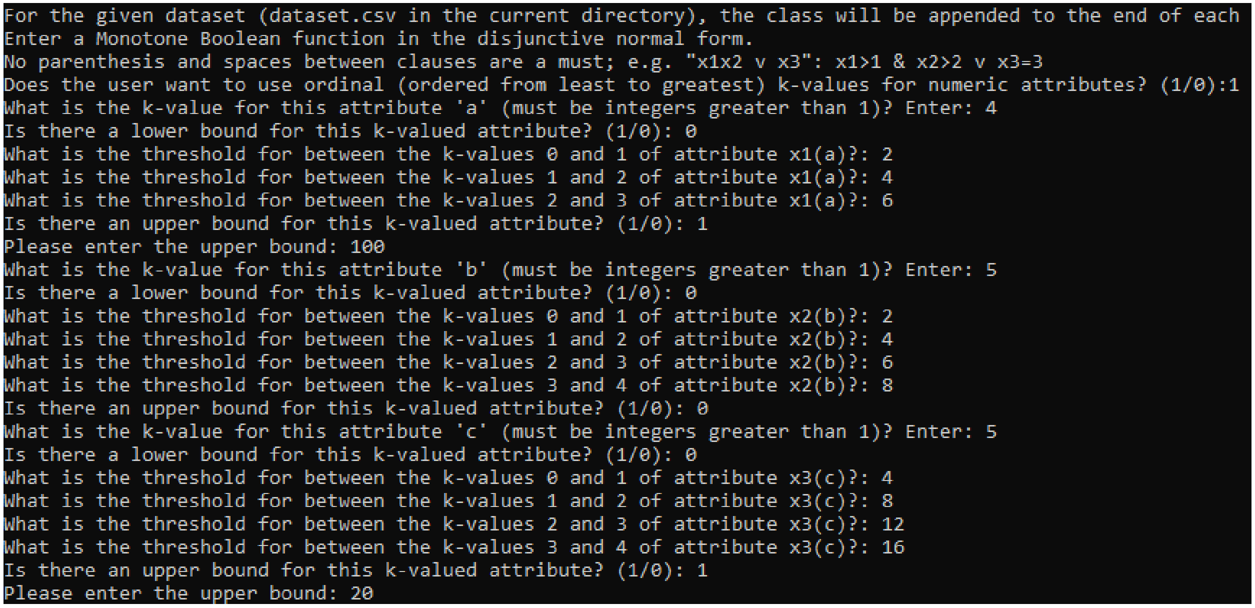
This program outputs a results file where the function values for each datapoint is given. The function that was given as input is printed as well as the k-value representation as well as the k-value for each attribute. For datasets where the k-value for that attribute is automatically calculated, then the output is as stated above (note: the k-value representation will be the same as the original data). For datasets where thresholds were needed, then the intervals that the thresholds represent are given as well. Note that the order of the datapoints in the output is by function value, where datapoints with the value of 1 are first.

Provided Files

There are several files besides the code files themselves. “kv_test.csv” is the input csv file that is a regular k-valued dataset for the program. “kv_test_numeric.csv” is similar, but all the values are numeric, so thresholds will need to be used. Currently, the program is set to the former file. Finally, there is a results file with example results for both these two datasets and several different k-value functions. There is also an executable provided, but it can only be used to run the file “kv_test.csv” since the file name for the dataset is declared in the header file.

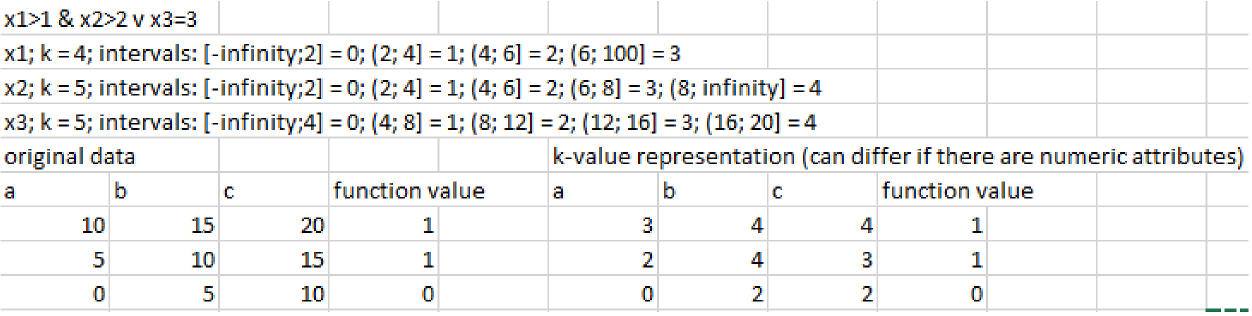
Examples

User input



```
For the given dataset (dataset.csv in the current directory), the class will be appended to the end of each
Enter a Monotone Boolean function in the disjunctive normal form.
No parenthesis and spaces between clauses are a must; e.g. "x1x2 v x3": x1>1 & x2>2 v x3=3
Does the user want to use ordinal (ordered from least to greatest) k-values for numeric attributes? (1/0):1
What is the k-value for this attribute 'a' (must be integers greater than 1)? Enter: 4
Is there a lower bound for this k-valued attribute? (1/0): 0
What is the threshold for between the k-values 0 and 1 of attribute x1(a)?: 2
What is the threshold for between the k-values 1 and 2 of attribute x1(a)?: 4
What is the threshold for between the k-values 2 and 3 of attribute x1(a)?: 6
Is there an upper bound for this k-valued attribute? (1/0): 1
Please enter the upper bound: 100
What is the k-value for this attribute 'b' (must be integers greater than 1)? Enter: 5
Is there a lower bound for this k-valued attribute? (1/0): 0
What is the threshold for between the k-values 0 and 1 of attribute x2(b)?: 2
What is the threshold for between the k-values 1 and 2 of attribute x2(b)?: 4
What is the threshold for between the k-values 2 and 3 of attribute x2(b)?: 6
What is the threshold for between the k-values 3 and 4 of attribute x2(b)?: 8
Is there an upper bound for this k-valued attribute? (1/0): 0
What is the k-value for this attribute 'c' (must be integers greater than 1)? Enter: 5
Is there a lower bound for this k-valued attribute? (1/0): 0
What is the threshold for between the k-values 0 and 1 of attribute x3(c)?: 4
What is the threshold for between the k-values 1 and 2 of attribute x3(c)?: 8
What is the threshold for between the k-values 2 and 3 of attribute x3(c)?: 12
What is the threshold for between the k-values 3 and 4 of attribute x3(c)?: 16
Is there an upper bound for this k-valued attribute? (1/0): 1
Please enter the upper bound: 20
```

Results



x1>1 & x2>2 v x3=3								
x1; k = 4; intervals: [-infinity;2] = 0; (2; 4] = 1; (4; 6] = 2; (6; 100] = 3								
x2; k = 5; intervals: [-infinity;2] = 0; (2; 4] = 1; (4; 6] = 2; (6; 8] = 3; (8; infinity] = 4								
x3; k = 5; intervals: [-infinity;4] = 0; (4; 8] = 1; (8; 12] = 2; (12; 16] = 3; (16; 20] = 4								
original data					k-value representation (can differ if there are numeric attributes)			
a	b	c	function value		a	b	c	function value
10	15	20	1		3	4	4	1
5	10	15	1		2	4	3	1
0	5	10	0		0	2	2	0