Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)

Handling of Numeric Ranges for Graph-Based Knowledge Discovery

Oscar E. Romero and **Jesus A. Gonzalez**
National Institute of Astrophysics, Optics and Electronics
Computer Science Department
Luis Enrique Erro 1, Mexico
oromero,jagonzalez@ccc.inaoep.mx

Lawrence B. Holder
Washington State University
The School of Electrical Engineering and Computer Science
Pullman, USA
holder@wsu.edu

Abstract

Nowadays, graph-based knowledge discovery algorithms do not consider numeric attributes (they are discarded in the preprocessing step, or they are treated as alphanumeric values with an exact matching criterion), with the limitation to work with domains that do not have this type of attribute or finding patterns without numeric attributes. In this work, we propose a new approach for the numerical attributes handling for graph-based learning algorithms. Our approach shows how graph-based learning approaches increase their accuracy for the classification task and its descriptive power when they are able to use both nominal and numerical attributes. This new approach was tested with the Subdue system in the mutagenesis and PTC domains showing an accuracy increase around 16% compared to Subdue when it does not use our numerical attributes handling algorithm.

In some research areas such as data mining and machine learning, the domain data representation is a fundamental aspect that determines in great measure the quality of the results of the discovery process. Depending on the domain, the Data Mining process analyzes a data collection (such as flat files, log files, relational databases, etc.) to discover patterns, relationships, rules, associations, or useful exceptions to be used for decision making processes and for the prediction of events and/or concept discovery. Graph based algorithms have been used for years to describe (in a natural way) flat, sequential, and structural domains with acceptable results (Gonzalez, Holder, and Cook 2002), (Ketkar, Holder, and Cook 2005). Some of these domains contain important numeric attributes (attributes with continuous values). Domains with continuous values are not appropriately manipulated by graph based knowledge discovery systems, although they can be appropriately represented. To the best of our knowledge there does not exist a graph based knowledge discovery algorithm that deals with continuous valued attributes. A solution proposed in the literature to approach this problem is the use of discretization techniques as a preprocessing or post-processing step but not at the knowledge discovery phase. However, we think that these techniques do not use all the available knowledge that can be taken advantage of during the processing phase. Adding this capacity to graph-based algorithms will allow us improving the work with numeric attributes and in this way we will be able to improve the classification accuracy for the classification task and the patterns descriptive power. We will then be able to enhance our results for structural domains containing numerical attributes.

Handling of Numerical Ranges

In this section, we describe the numerical ranges generation algorithm (based on frequency histograms), which calculates distances using any of seven measures (the distance between the values is calculated by any of seven methods). The distances are: a modification to the Tanimoto distance, a modification of the Euclidean distance, a modification to the Manhattan distance, a modification to the Correlation distance, a modification to the Canberra distance, and two new distance measures that we propose. Our algorithm can be seen in figure 1. The algorithm shown in figure 1 works as follows. The general function (GenerateRange) receives the data set and the number of examples in it. In the first step we sort the numerical attribute in ascending way (Sort function). Next, we create a frequency histogram of the ordered data. Then, we create an initial table of ranges with four fields corresponding to the center of the range, its frequency, and its low and high limits. In this initial ranges table, the center, the low and high limits contain the same value take from the frequencies histogram (GenerateHistogram function). After

```
GenerateRange (data, N)
    Sort (data)
    for i = 1 to 7
        new histo
        SetInitial (data, N, histo)
        GenerateHistogram (histo)
        distance = TypeofDistance(i, Average(histo), Center(histo), histo)
        threshold = distance + Minimal (histo)
        TypeofGroup(i, threshold, histo)
        rangetable[i] = histo
    return rangetable
```

Figure 1: Numerical Ranges Generation Algorithm

that, we calculate the minimum distance between any two consecutive row, using their center fields (Minimal function), and we calculate an average of all the center fields of the frequency histogram (Average function). After that, we calculate the centroid for the center fields of the table of ranges, which corresponds to the element of the frequency histogram closest to the value of the average (Center function). We now calculate a type of distance that is used to calculate the grouping threshold (to decide which values are grouped to create a numerical range). In the next step we calculate a grouping threshold, which is the sum of the minimum distance plus the distance. After that, we iteratively group the elements of the ranges table until the table of ranges does not suffer any modification (TypeofGroup function). At this step we have obtained a final table of ranges.

Results

In this work we experimented with different data representations of numerical ranges for two structural databases. These databases are: Mutagenesis and PTC (The Predictive Toxicology Challenge for 2000-2001). Both databases were created by the National Institute of Environmental Health Sciences (NIEHS). In order to show how Subdue can be enhanced when adding the capability to deal with numeric attributes, we performed our first experiment, in which we did not give any special treatment to each numerical attribute. With our second test we show that Subdue is able to find interesting patterns containing numerical values when we generated numerical ranges (also improving its classification accuracy). We included our numerical ranges information to our graph-based representations and executed a 3-fold cross-validation with Subdue. We can see the results for the mutagenesis database in table 1 and for the PTC database in table 2. Table 1 shows the results obtained for the mutagenesis

Type of Graph-Based Data Representation		Regression	
		Unfriendly	**Friendly**
Without Rings	Without Ranges	58.54%	69.39%
	With Ranges	82.85%	89.00%
With Rings	Without Ranges	61.54%	72.13%
	With Ranges	84.36%	90.80%

Table 1: Accuracy achieved for the Mutagenesis database using a 3-fold-cross-validation.

database with and without the use of numerical ranges for both data representations (with and without rings). The behavior of the results for both representations is stable. We think that by adding rings to representation "I" to create representation "II" we obtained better accuracy results and more descriptive models (with structural information about rings). The input graph of representation "II" is larger than the one created for representation "I". This means that the search space for representation "II" is larger than the one for representation "I" and we need to increase Subdue's parameters in order to find a better model in terms of classification accuracy. However, we obtain a 21% increment when using numerical attributes with both data representations, which means that providing Subdue the capability to handle numerical ranges makes it able to find better models. Table 2 shows the results obtained with Subdue (with

Type of Graph-Based Data Representation		PTC			
		MM	**FM**	**MR**	**FR**
Without Rings	Without Ranges	66%	62%	54%	58%
	With Ranges	74%	70%	62%	66%
With Rings	Without Ranges	69%	65%	57%	61%
	With Ranges	78%	74%	76%	80%

Table 2: Accuracy achieved for the PTC database using a 3-fold-cross-validation.

and without ranges) for the PTC database. In this table it is possible to observe that, on average, the classification accuracy increased 11% when we use our proposed method to handle numerical ranges for both data representations (with and without rings). This accuracy increment is not as high as we expected it to be, but as in the previous table, it is due to the execution of Subdue with limited parameters. We can also see in the table that the classification accuracy for all the subsets of both databases (PTC and Mutagenesis databases) is stable, which differs from the results reported in related works.

Conclusion and Future Work

The main contribution of this work considers a guideline for the creation of a graph-based representation for mixed data types (continuous and nominal) and also the creation of an algorithm for the manipulation of these graphs with numerical ranges for the data mining task (classification and discovery). For our future work we will test different domains to enrich the results of our approach. We will also include temporal information with numerical values. After we have collected this data, we will be able to perform a spatial and temporal data mining process. Finally we will compare our results with Subdue against other algorithms that can deal with structural representations such as inductive logic programming systems.

Acknowledgements

The first author acknowledges Conacyt for the support provided in my doctoral studies with the scholarship number 86997.

References

Gonzalez, J. A.; Holder, L. B.; and Cook, D. J. 2002. Experimental comparison of graph-based relational concept learning with inductive logic programming systems. In *Lecture Notes in Artificial Intelligence*, volume 2583, 84–99. Springer Verlag.

Ketkar, N. S.; Holder, L. B.; and Cook, D. J. 2005. Comparison of graph-based and logic-based multi-relational data mining. *SIGKDD Explor. Newsl.* 7(2):64–71.