Handling of Numeric Ranges with the Subdue System

Oscar E. Romero A. and **Jesus A. Gonzalez B.**
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

Abstract

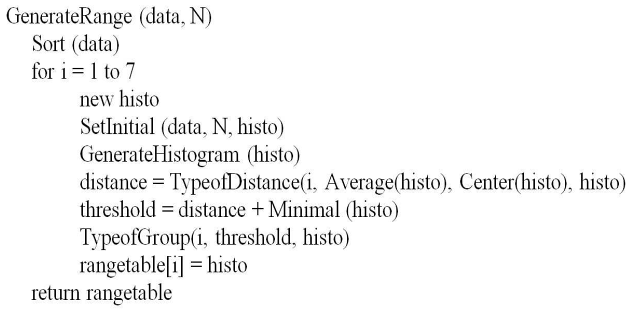
Graph-based knowledge discovery has become a powerful tool in the machine learning and data mining areas. It provides a flexible and natural data representation to describe real world domains. In this research work we present a novel algorithm for graph-based approaches to deal with numerical attributes during the data processing phase implemented in the Subdue system. Our experimental results show that the use of numerical attributes increased classification accuracy in the Mutagenesis and PTC domains in 22% compared to the Subdue system when it does not use our numerical attributes handling approach. Our method also outperforms other author's results for the same domains, around 7% for the Mutagenesis domain and around 17% for the PTC domain.

Recently, artificial intelligence techniques have become tools to solve real world problems and the need to represent real world domains has increased. Many interesting real world domains have an inherently structured character for which we need to find data representations to effectively apply AI algorithms to them. Data domains maybe classified as flat, sequential, and structural according to their properties. Some of these domains contain important numeric attributes. 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 at the time of publishing this work a graph-based knowledge discovery algorithm that deals with continuous values attributes does not exist. A solution proposed in the literature to approach this problem is the use of discretization techniques as a pre-processing or post-processing step but not during the knowledge discovery phase. Adding this capacity to graph-based algorithms will allow us to improve the work with numeric attributes; in this way we will be able to improve the classification accuracy for the classification task and the patterns descriptive power.

Subdue

Subdue (Cook and Holder 1994) is a graph-based knowledge discovery system that has been successfully used in different structural domains such as chemical compounds and DNA. The input to Subdue is a graph that represents objects and their attributes with labeled vertices and edges, where vertex labels give names to objects or their attributes. Relations between objects and relations among objects with their attributes are represented by labeled edges, where the label determines the name of the relation. This allows working with any domain that can be represented with graphs. Subdue's evaluation model decides which patterns are going to be chosen as important. For this, Subdue implements two evaluation criteria. The first evaluation method is called "Minimum Encoding", which is a variant of the minimum description length principle (MDL) introduced by Rissanen. The second method, "Graph Size", chooses the best substructure according to how well it compresses the graph in terms of its size in number of vertices and edges. Subdue is also able to perform the concept learning task and this version of the algorithm is known as SubdueCL (Gonzalez, Holder, and Cook 2002). SubdueCL, such as CProgol, follows a set covering approach. It is important to mention that neither Subdue nor SubdueCL deals with numerical attributes during the processing phase.

Algorithm



```
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

In this section, we show the numerical ranges generation algorithm, which calculates distances using any of seven measures. Our algorithm can be seen in figure 1.

Results

In this section we present the most significant results for the experiments with the Mutagenesis and PTC datasets. In order to show how Subdue can be enhanced when adding it the capability to deal with numeric attributes, we performed our first test, in which we did not give it any special treatment to any numerical attribute. With the second test, we show that the "Subdue" system is able to find interesting patterns containing numerical values when we generate numerical ranges.

Table 1: Accuracy achieved for the PTC database.

Algorithms	**Type of Data Representation**		**PTC**			
			FM	**FR**	**MM**	**MR**
Subdue	Without Rings	Without Ranges	62%	58%	66%	54%
		With Ranges	70%	70%	74%	64%
	With Rings	Without Ranges	65%	61%	69%	57%
		With Ranges	**74**%	**83**%	**78**%	**76**%
CProgol	Without Rings	Without Ranges	54%	52%	53%	55%
MCS	(Maji and Mehta 2006)		72%	79%	77%	75%
MCES	(Maji and Mehta 2006)		73%	79%	77%	73%
Marginalized	(Kashima, Tsuda, and Inokuchi 2003)		63%	67%	64%	63%
Tanimoto	(Ralaivola et al. 2005)		64%	67%	66%	66%
Minimax	(Ralaivola et al. 2005)		65%	66%	64%	65%

We can see the results for the Mutagenesis and PTC databases in table 2. The behavior of the results for both representations is similar. We think that by adding rings to representation "**I**" (which uses numerical data and some basic relations between its attributes) to create representation "**II**" (we add relations based on the rings components) we should have obtained better accuracy results and more descriptive models (with structural information about rings). Analyzing our results we can see that when we add structural data to representation "**I**" to obtain representation "**II**", accuracy increases by 15% (on average) with respect to the accuracies obtained without using rings. The table shows the results obtained when we use CProgol with representation "**II**". The results that we obtained with CProgol are slightly inferior (around 3 to 4%) than those reported in the literature. This might have happened because the background knowledge (or the parameters setting) used in those other works could be different than those used by us. We should also consider that the "Subdue" system does not use background knowledge and that we use the "Subdue" system with limited parameters. We compared the results of the previous table (Mutagenesis and PTC databases) with the results of other authors (as shown in the same table, last rows). As we can see, the numerical ranges handling (using numerical and structural data at the same time) that we used with the Graph-Based system: "Subdue", increased Subdue's accuracy with respect to other algorithms. In these results we can see that our approach obtains an increase of almost 30% for both, the PTC and the Mutagenesis databases.

Table 2: Accuracy achieved for the Mutagenesis database.

Algorithms	**Type of Data Representation**		**Mutagenesis**	
			42	**188**
Subdue	Without Rings	Without Ranges	59%	69%
		With Ranges	87%	89%
	With Rings	Without Ranges	62%	72%
		With Ranges	**84**%	**91**%
CProgol	Without Rings	Without Ranges	67%	80%
MCS	(Maji and Mehta 2006)			84%
MCES	(Maji and Mehta 2006)			87%
Marginalized	(Kashima, Tsuda, and Inokuchi 2003)			90%
Tanimoto	(Ralaivola et al. 2005)			88%
Minimax	(Ralaivola et al. 2005)			86%

Conclusion and Future Work

As our main conclusion we have that the different ways to generate numerical ranges helped our method to find significant patterns at different levels of abstraction (through our different graph-based data representations) using the Subdue system. We can also note that with our approach to create dynamic numerical ranges we were able to obtain richer descriptive patterns that were also able to obtain better accuracy for the classification task. This evidence shows that our novel approach for numerical ranges handling for graph-based approaches is promising. In this way, and as our main contribution, we created a graph-based approach able to deal with mixed data types (nominal and continuous) for the concept description and classification tasks. In our future work we will experiment with different real world domains to take advantage of our approach. With our new approach we will be able to perform the data mining task with spatio-temporal domains.

Acknowledgements

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

References

Cook, D. J., and Holder, L. B. 1994. Substructure discovery using minimum description length and background knowledge. In *Journal of Artificial Intelligence Research*, volume 1, 231–255.

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.

Kashima, H.; Tsuda, K.; and Inokuchi, A. 2003. Marginalized kernels between labeled graphs. In *Proceedings of the Twentieth International Conference on Machine Learning*, 321–328. AAAI Press.

Maji, S., and Mehta, S. 2006. A netflow distance between labeled graphs: Application in chemoinformatics. *Undergraduate Thesis. Department of Computer Science of Indian Institute of Technology, Kanpur* 1(1):1–6.

Ralaivola, L.; Swamidass, S. J.; Saigo, H.; and Baldi, P. 2005. Graph kernels for chemical informatics. *Neural Netw.* 18(8):1093–1110.