ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

k-NN behavior with set-valued attributes

Mabel González, Yanet Rodríguez and Carlos Morell.

Universidad Central de Las Villas - Computer Science Department
Carretera a Camajuaní km 5 ½, Santa Clara, Villa Clara - Cuba
{mabelc,yrsarabia,corellp}@uclv.edu.cu

Abstract. This paper addresses the problem of dealing with set-valued attributes in the lazy learning context. This type of attribute is present in various domains, yet the instance-based learning tools do not provide a representation for them. To solve this problem, we present a proposal for the treatment of the sets in the context of the *k*-NN algorithm through an extension to HEOM distance. Experiments using various data sets show the feasibility of this option.

1 Introduction

The set term plays a fundamental role in the development of the modern mathematics. We can define a set, like a collection of objects with well defined characteristics that includes it on a certain group. A set exists if have the following properties[1]:

- The collection of elements should be well defined.
- The elements belonging to a set can not be repeated.
- The order of the elements is not important.

Set-valued attributes appear in a natural way for several real world problems. A clear example is present in medical diagnosis. The diagnosis is obtained starting from a set of symptoms and signs; these can be present simultaneously in the same person. On the other hand, in text categorization, a document is treated as a set of keywords; the domain of words doesn't have a fixed number and is usually large. We only will consider finite sets in this paper.

The way in which training examples are represented is of critical importance for the concept learning system. In most implemented concept learning systems an example is represented by a fixed length vector, the components of which are called attributes or features. Most practical machine learning systems [2-5] accommodate just two levels of measurement: nominal and ordinal [5]. Nominal attributes are sometimes called categorical, enumerated, or discrete.

There are two typical solutions to deal with data that are naturally represented with set-valued features [6]. The former is to repeat each instance according to the number of existing distinct combinations. Each instance will contain a different combination of possible values of each set-valued attribute. The previous approach would cause an increase of the training instances number. The later is to create a Boolean vector for each set-valued attribute. The size of the vector is equal to the cardinality of the domain to which the set-valued attribute belongs. The original sets are mapped to vectors by setting true if the corresponding value appears in the set and setting the remainder to false. This second way to deal with the problem has as a disadvantage, instances with larger numbers of attributes. Instances obtained using this method are, in many cases, unmanageable like the categorization text problem previously mentioned. The two solutions discussed above have another weakness; the possible loss of information caused by the decomposition of the original sets.

A more current approach is to consider the treatment of sets as another data type where the challenge is how to handle the set-valued attribute in the context of each algorithm. Recently, several variants of this approach have been proposed, such as: an approach for modifying rule induction algorithms to learn from sets [7] and a k-Nearest Neighbor (k-NN) learning algorithm, IBPL that utilizes the Value Difference Metric [8], and accepts instances containing set-valued attributes [9].

This paper proposes the use of Jaccard distance function between sets in the contexts of k nearest neighbor classification to deal with set-valued attributes. That is, to modify a learning schema to deal with a new data type. This approach proved to be successful and we will show, empirically, a comparative study between both approaches using several datasets obtained from UCIMLR. All tests have been performed in the Waikato Environment for Knowledge Analysis (WEKA), which includes working implementations of several machine learning methods.

2 Handling set-valued attributes

The machine learning concept includes any computer program that improves its performances at some task through experience [10]. Experience is provided by training examples belonging to a problem domain. On the other hand k-NN algorithms also known as Instance-Based learning algorithms [11], fall into the category of lazy learning algorithms. An important point in this type of algorithm is the choice of the similarity function used to compare instances.

One way to handle applications with both continuous and nominal attributes is to use the distance function HEOM [12], which uses different attribute distance functions on different kinds of attributes. If at least one of the two values is a missing value (?), the maximum distance is applied. This distance uses the overlap metric for nominal attributes and normalized Euclidean distance for linear attributes, where l_a is the lower bound of attribute and u_a is the upper bound. Distance definition (1) returns a value which is in the range 0...1, where one means maximum distance and 0 minimum. Finally the global distance between instances is based on the distances among attributes values.

$$d_{HEOM}(x,y) = \begin{cases} 1 \text{ if } \mathrm{x} = ? \text{ or } \mathrm{y} = ? \\ \delta_{\mathrm{overlap}}(x,y) & \text{symbolic attribute} \\ \delta_{norm1}(x,y) & \text{linear attribute} \end{cases} \tag{1}$$

$$\delta_{norm1}(x,y) = \frac{|x-y|}{l_a - u_a} \tag{2}$$

$$\delta_{overlap}(x,y) = \begin{cases} 0 & \text{if } \mathrm{x} = \mathrm{y} \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

For calculating the distance between two sets we use Jaccard distance (5). The Jaccard distance, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient [13] (4) and is obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union.

$$S_{Jaccard}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{4}$$

$$\delta_{Jaccard}(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} \tag{5}$$

Finally the above equation that allows for the calculation of the distance between sets is added to HEOM distance (6). Jaccard distance returns a value which is in the range 0...1 like the original HEOM distance (1). Now the new HEOM (6) accepts instances containing set-valued attributes.

$$d_{HEOM}(x,y) = \begin{cases} 1 \text{ if } \mathrm{x} = ? \text{ or } \mathrm{y} = ? \\ \delta_{\mathrm{overlap}}(x,y) \quad \text{symbolic attribute} \\ \delta_{norm1}(x,y) \quad \text{linear attribute} \\ \delta_{Jaccard}(X,Y) \ \text{set - valued attribute} \end{cases} \tag{6}$$

3 Empirical Comparisons and Discussion

We made an extension to WEKA[*] to include the new data type. Furthermore we modified the HEOM distance source according to the approach proposed. With the objective of showing the feasibility of our proposition, experiments were made in order to compare the variant based on set-valued attribute and the decomposition in binary attributes. The performance of the proposed distance was evaluated on six data sets where the class attribute is nominal: leptospirosis data, pneumonia data, pediatric data, hepatitis data, primary-tumor data, SPECT data, voting data and flags data. These last three data sets can be obtained from the machine-learning database at UCI Machine Learning Repository.

The primary tumor domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Class attribute is location of tumor. Pneumonia data corresponds to a dataset that allows the pneumonia diagnosis, the same occurs with the Leptospirosis data. Pediatric [14] can predict the gravity of a pediatric patient on arrival at hospital. For more information about data sets and set-valued attributes defined see Table 1.

[*] WEKA is a Java-written open source. It is available at http://www.cs.waikato.ac.nz/˜ml/weka/ under the GNU General Public License.

Datasets	Nominal attributes	Numeric attributes	Set-valued attributes	# of classes	Domains
Primary Tumor	5	0	1 (organs)	22	{bone-marrow, pleura, peritoneum, liver, brain, skin, supraclavicular, axillar, lung, bone, mediastinum, abdominal, neck}
Hepatitis	4	5	1 (symptoms)	2	{fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, steroid, antivirals}
Flags	15	2	2 (colors, forms)	6 10 7†	{red, green, blue, gold, white, black, orange} {crescent, triangle, icon, animate, text}
SPECT	1	0	1 (patterns)	2	{F1, F2, F3, F4, F5, F6, F7, F8,…,F22}
Voting	1	0	1 (votes)	2	{handicapped-infants, water-project-cost-sharing,…, export-administration-act-south-africa}

Leptospirosis	2	1	7 (symptoms, animals, water, housing, bath, tests, activities)	2	{symptoms$_1$,…, symptoms$_{26}$}
					{animals$_1$,…,animals$_{15}$ }
					{water$_1$,…, water$_{12}$}
					{housing$_1$,…, housing$_6$}
					{bath$_1$,…, bath$_7$}
					{tests$_1$,…, tests$_{10}$}
					{activities$_1$,…, activities$_{14}$}

Table 1: Attribute Characteristics.

Datasets were classified with the *k*-NN algorithm using the best *k* in each case. The accuracy is obtained by means of a cross-validation process with ten folds. The best k was found using leave one out cross validation over training set. The average accuracy for each dataset over all trials is used to rank the classifiers employed during

[†] Number of classes depends of the target attribute.

the comparison. Table 2 shows the average accuracy for each dataset and the *k* values used.

Dataset	Binary	k	Set-Valued	k
Hepatitis	83.87	5	**85.16**	7
Primary-Tumor	**47.49**	14	43.66	14
SPECT	78.65	7	**81.65**	7
Pediatric	88.65	1	**89.21**	5
Pneumonia	67	5	**70**	5
Flags-landmass	51.51	5	**62.37**	14
Flags-language	46.39	14	**54.12**	8
Flags-religion	57.22	5	**61.34**	7
Voting	92.64	4	**93.33**	1
Leptospirosis	92.16	2	**93.14**	8

Table 2: Comparison between the two approaches.

The non parametric the Wilcoxon signed-rank test was used to compare the model proposed with the usual approach. Table 3 shows the results of the Wilcoxon signed-rank test. The significance is based in 10000 sampled tables with Monte Carlo simulation techniques and a 99% confidence interval. The test found significant differences (see Table 4) in favor to the variant based on set-valued attribute.

		N	Mean Rank	Sum of Ranks
SetValued - Binary	Negative Ranks	1(a)	7.00	7.00
	Positive Ranks	9(b)	5.33	48.00
	Ties	0(c)		
	Total	10		

Table 3: Results of the Wilcoxon signed-rank test (a: SetValued < Binary, b: SetValued > Binary and c: SetValued = Binary).

	SetValued – Binary
Z	-2.091(a)
Asymp. Sig. (2-tailed)	.037

Table 4: Test statistics (a: Based on negative ranks. b: Wilcoxon Signed Ranks Test).

The above results show how the modeling of attributes using set-valued attributes improve the k-NN performance with respect to the variant based on decomposition into binary attributes.

Conclusions and further work

The proper treatment of sets must be taken into account in problem modeling. Real world problems that can be naturally expressed with set-valued features are not as rare as one might think. Decompose attributes in problems that merits a set-valued representation results in information lost that subsequently affects the classification.

The modification to the HEOM distance proposed, allows the treatment of sets like another data type. Jaccard distance was used to calculate similarity between sets. Principal advantages of this approach are its simplicity, efficiency and a problem modeling way more natural.

Currently we are evaluating the extension of other machine learning algorithms to make possible the use of set-valued attributes. From the behavior of different approach algorithms, we will arrive at more definitive conclusions about the impact of the correct treatment of sets in the machine learning context.

References

[1] Devlin, K.J., The Joy of Sets. 1993, Springer-Verlag. p. 1-7.

[2] Alcalá-Fdez, J., et al., KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing, 2009. 13(3): p. 307-318.

[3] Demsar, J., B. Zupan, and G. Leban, Orange: From Experimental Machine Learning to Interactive Data Mining. 2004, Faculty of Computer and Information Science, University of Ljubljana.

[4] Mierswa, I., et al. YALE: Rapid Prototyping for Complex Data Mining Tasks. in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06). 2006.

[5] Witten, I. and D. Frank, Data Mining Practical Machine Learning Tools 2. 2005: Elsevier.

[6] Payne, T.R., Dimensionality Reduction for Agent-Based Learning. 1996.

[7] Cohen, W.W. Learning Trees and Rules with Set-valued Features. in Proceedings of the Thirteenth National Conference on Artifcial Intelligence. (AAAI-96). 1996.

[8] Stanfill, C. and D. Waltz, Toward memory-based reasoning. Communications of the ACM, 1986. 29(12): p. 1213-1238.

[9] Payne, T.R., Dimensionality Reduction and Representation for Nearest Neighbour Learning. 1999, University of Aberdeen.

[10] Mitchell, T., Machine Learning. 1997: McGraw Hill.

[11] Aha, D. and D. Kibler, Instance-based learning algorithms. Machine Learning, 1991. 6: p. 37-66.

[12] Wilson, D.R. and T.R. Martinez, Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research, 1997. 6(1): p. 1-34.

[13] Tan, P.-N., M. Steinbach, and V. Kumar, Introduction to Data Mining. 2005.

[14] Sarabia, Y.R., et al., Prediction of Pediatric Risk Using a Hybrid Model Based on Soft Computing Techniques, in MICAI 2008: Advances in Artificial Intelligence. 2008, Springer Berlin / Heidelberg. p. 472-481.