Applications of Rough Sets Theory in Data Preprocessing for Knowledge Discovery

Frida Coaquira and Edgar Acuña

***Abstract*— Data preprocessing is a step of the Knowledge discovery in databases (KDD) process that reduces the complexity of the data and offers better conditions to subsequent analysis. Rough sets theory, where sets are approximated using elementary sets, is a different approach for developing methods for the data preprocessing process. In this paper Rough sets theory is applied to three preprocessing steps: Discretization, Feature selection, and instance selection. The new methods proposed in this paper have been tested on eight datasets widely used in the KDD community.**

***Index Terms*— Rough sets, feature selection, instance selection, Knowledge Discovery.**

I. Introduction

Rough sets theory was introduced by Z. Pawlak (1982) as a mathematical tool for data analysis. It does not need external parameter to analyze and make conclusions about the datasets. Rough sets offer many opportunities for developing many Knowledge Discovery methods using partition properties and discernability matrix [18], [19], [21], [22]. Rough sets have many applications in KDD among them, feature selection, data reduction, and discretization [1], [5], [6], [14], [17]-[20]. Rough sets can be used to find subsets of relevant (indispensable) features [5], [7]. Combining rough sets theory with a known classifier yields a wrapper feature selection method since it uses the class label information to create the indiscernability relation. It provides a mathematical tool that can be used to find out all possible feature subsets [13], [14]. In Feature selection problem, the purpose of using Rough sets is to find the indispensable features. The principal idea is to recognize the dispensable and indispensable features, using the discernibility matrix [11], [12], [14], [18].

Manuscript received July 13, 2007. This work was supported in part by ONR under Grant N00014060555.

F. Coaquira is with the Department of Mathematical Science, University of Puerto Rico at Mayaguez, Mayaguez, PR 00680 USA (e-mail: frida_cn@ math.uprm.edu)

E. Acuna is with the Department of Mathematical Science, University of Puerto Rico at Mayaguez, Mayaguez, PR 00680 (corresponding author, e-mail: edgar@math.uprm.edu).

II. BACKROUND ON ROUGH SET

Let $T = (U, A, C, D)$ be a decision system data, where U is a non-empty finite set called the universe, A is a set of features, C and D are subsets of A, named the conditional and decisional attributes subsets respectively.

Definition 1. Let $R \subseteq C$ and $X \subseteq U$, the R-lower approximation set of X, is the set of all elements of U which can be with certainty classified as elements of X.

$$\underline{R}X = \cup\{Y \in U / R : Y \subseteq X\}$$

According to this definition, we can see that R-Lower approximation is a subset of X,

thus $\underline{R}X \subseteq X$.

Definition 2. The R-upper approximation set of X is the set of all element of U, that can belong possibly to the subset of interest X.

$$\overline{R}X = \cup\{Y \in U / R : Y \cap X \neq \phi\}$$

Note that X is a subset of the R-upper approximation set, thus $X \subseteq \overline{R}X$.

Definition 3. The Boundary region of a set X is the collection of elementary sets defined by

$$BN(X) = \overline{R}X - \underline{R}X$$

These sets are included in R-Upper but not in R-Lower approximations.

Definition 4. A subset defined through its lower and upper approximations is called a Rough set. That is, when the boundary region is a non-empty set ($\overline{R}X \neq \underline{R}X$).

Definition 5. Let $T = (U, A, C, D)$ be a decision table, the Dependency Coefficient between the conditional attribute C, and the decision attribute D is given by

$$\gamma(A, D) = \frac{card(POS(C, D))}{card(U)}$$

where, *card* denotes cardinality of a set.

The dependency coefficient expresses the proportion of the objects correctly classified with respect the total,

considering the set of conditional features. The dependency between attributes in a data set is an important issue in data analysis. A decisional attribute depends on the set of conditional features if all values of decisional feature D are uniquely determined by values of conditional attributes. i.e. there exists a dependency between values of decisional and conditional features [21],[22].

III. Discretization

Discretization is the process to transform continuous features into qualitative features. Firstly, continuous feature values are divided into subintervals. Then, each interval is mapped to a discrete symbol (categorical, nominal or symbolic) [4]. These discrete symbols are used as new values of the original features. A cut point is a real value c, within the range of a continuous feature, that partitions the interval [a, b] is partitioned into two subintervals [a, c] and (c, b]. A continuous feature could be partitioned into many subintervals. A continuous feature with many cut points can make the learning process longer, while a very low number of cut points may affect the predictive accuracy negatively [15]. A number *m* could be considered as an upper bound for the number of cut points. In practice, *m* is set to be much less than the number of instances, assuming there is no repetition of continuous value for a feature [25]-[26].

Rough sets theory can be applied to compute a measure considering partitioning generated by these cut points and the decisional feature in order to obtain a better set of cut points. We set *m* as the number of intervals given by the Scott's formula to determine the bins of a histogram. The proposed algorithm is as follows:

Input: The original dataset D , and *m* the maximum number of intervals to be considered

For each continuous feature v_i of Data

For j in 1:m (m is nclass.scott(v_i))

Calculate the partition considering j equal width intervals

Evaluate each partition using an association measure based on Rough sets

$$\gamma = \frac{Pos(v_i / d)}{n}$$

Stopping criteria: Select the optimal number of partition p

Figure 1. Discretization algorithm based on Rough sets

IV. Feature Selection

The problem of feature selection consists in the search of *d* features from a given set of *m* ($d<m$) features, that in general leads to the smallest classification error rate. Feature selection methods determine an appropriate feature subset such that the classification error is optimal. The chosen features permit that pattern vectors belonging to different categories occupy compact and disjoint regions in an *m-dimensional* feature space [8]-[10],[13],[20],[27].

Dimension reduction is needed when the dataset has a large number of features. Classification and regression algorithms could present problems in their general behavior when redundant features are considered. This is a main reason for many investigators to search for different methods to detect these features. In reducing the number of features it is expected that the ones that are redundant and irrelevant will be deleted.

There are two main reasons to keep the dimensionality of the features as small as possible: cost minimization and classification accuracy. Cost minimization is achieved because after feature selection the classifiers will be computed faster and will use less memory [6]. A careful choice of the features is needed since a bad reduction may lead to a loss in the discrimination power and thereby a decrease in the accuracy of the resulting classifier [23],[24],[27],[29].

Feature selection by ranking according to dependency.

The best features can be found by calculating the dependency measure between any conditional feature and the decisional feature. After that, a ranking of features can be done. A basic filter algorithm to perform feature selection based on rough sets is shown in Fig. 2. This method calculates the dependency between every conditional feature considering the decisional feature, after ranking only the features with higher dependency values are included in the final subset of best features.

Input: Set of conditional and decisional features C, D.
Initialize the best subset of features as the empty set

i. For i in 1:number of conditional features
Apply some evaluation measure based on dependency of Rough sets.
End for
ii. Order the features according to dependency measure
iii. Select only the features with high dependency measure.
Output: A subset of features.

Figure 2**.** Algorithm for feature selection based on Rough sets.

V. Instance Selection

An instance or case is a collection of values considering all features for a given observation. Case selection in a dataset is carried out to obtain an appropriate subset of instances to perform a KDD task [16],[28].

In a dataset there are some instances that are

representative of elementary blocks of instances, then extracting a subset of interesting instances is related to set weights to each elementary sets obtained through Rough sets theory. An instance in a dataset is inconsistent when it has all their feature values similar to other instance but both of them lie in different classes. These instances should be analyzed carefully. Elementary sets formed using the set of conditional features help to identify the weight class where there should be inconsistent instances. Instances selection reduces the computation time of executing some KDD tasks since a new smaller dataset is obtained.
Instance selection is a way to reduce the size of the dataset, when it contains so many instances to be analyzed. A good sampling will reduce the computational complexity of data mining algorithm.

Instance selection algorithm using Rough sets theory

Our algorithm combines to criterion considered by Cano[3], firstly we discarded the inconsistent data. After that, we select a random sample from the positive region.

Input: The original dataset, which it might some continuous conditional feature, and the 100p percentage of instances to be sampled from the positive region.

1. Discretize continuous features

2. Find out the elementary sets, making partitions according to conditional and decisional features.

3. Determine the positive region to eliminate the inconsistent cases.

4. Select the labels of the 100p% instances within the positive region and save in a list L.

5. Extract cases from the original dataset according to L.

Output: The set of cases to be selected.

Figure 3. Algorithm for instance selection using Rough sets.

VI. RESULTS

Dependency between the conditional feature and the decisional feature is the most important item in the study. The results was calculate used R program and the Dprep library[2].
Fig. 4 represents the dependency between each conditional feature and the decisional feature in the Diabetes dataset.

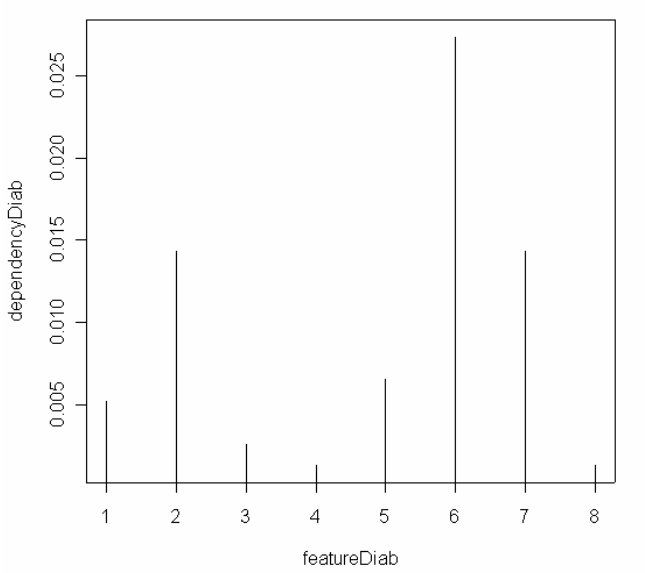

Figure 4. Dependency for each feature on Diabetes dataset.

Table I contains the cut points obtained for five discretization methods: The Entropy method, the 1R method, the ChiMerge method, the equal width using the Scott's formula, and the Rough set algorithm. The experimentation was carried on eight well known datasets.

Table II shows misclassification error rates using the LDA classifier and the discretized data. In the entropy-based discretization some features could not be used for the LDA algorithm because they do not have variability, since these features have only one value after the discretization process. These features appear among parentheses.

Table III contains the subset of features selected using a rough set criteria, after considering previously four different types of discretization : Rough Set method, 1R method, Equal width, and ChiMerge. The table also includes the best subsets of features using feature sequential selection (SFS) methods along with two classifiers: LDA and KNN.

Tables IV and V show the misclassification error rates for the LDA and KNN classifier, respectively, after feature selection is performed using only features from table III..

Finally, the instance selection algorithm was applied on the eight datasets, and then the KNN classifier was evaluated using only the selected cases. Table VI shows the average misclassification error rate on ten random samples (test samples) representing a 30% of the original data To evaluate the effect of the algorithm, we compute the misclassification error after and before instance selection. We have considered only data discretized by rough sets and Chi-Merge methods.

Table I. Comparison of the number of cutpoints for feature using five discretization methods.

Dataset	Rough Method	Equal width Bound Scott	Entropy Method	1R method	Chi-merge
Iris	6 4 4 4	7 9 6 5	3 3 3 3	3 3 3 3	7 5 4 4
Glass	9 14 2 10 12 10	13 14 6 11 13 13	3 2 2 3 1 4	6 9 8 3 9 4	15 7 7 9 8 9
Diabetes	5 13 16 8 19 21 16 6	14 17 17 17 20 23 19 14	2 4 1 1 3 2 2 2	8 6 6 10 14 16 8 15	5 14 4 9 41 45 80 9
Heartc (1,4,5,8,10,14)	8 6 17 6 10	11 12 17 11 11	2 1 1 2 2	5 8 9 2 4	6 6 32 18 8
Ionosfera	7 10 2 2 8 2 6 5 2 8 2 9 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	9 10 8 9 9 8 8 9 8 9 7 9 7 9 7 9 7 8 7 8 7 8 7 8 8 8 7 8 8 8 8 9	4 5 4 6 3 5 5 4 5 5 6 4 5 5 6 3 6 3 5 5 5 3 5 3 3 3 5 3 5 3 5 5	3 5 5 5 6 5 5 5 5 5 7 5 7 6 6 5 7 5 7 5 7 5 5 5 2 3 6 6 6 5 5 5	17 50 23 31 23 24 40 33 24 32 39 45 35 44 42 51 41 33 58 34 30 35 34 28 47 38 37 32 42 26 51
Crx (2,3,6,8,14, 15,16)	2 13 2 2 17 35	14 14 8 21 30 48	2 2 3 2 2 2	12 11 15 10 14 9	10 6 6 4 11 7
Vehicle	10 2 2 14 23 20 9 9 3 13 15 13 9 18 2 2 2 2	16 12 13 19 32 32 13 13 13 14 17 13 14 28 13 13 14 11	5 4 4 3 4 4 5 5 5 5 4 7 4 3 2 2 5 2	18 26 16 20 30 18 20 19 15 21 24 20 27 28 29 33 22 22	6 5 15 11 6 4 9 10 6 6 13 9 4 5 3 3 5 7
German (2,5,13,21)	6 13 2	17 19 15	2 2 1	9 6 5	10 27 2

Table II. Misclassification error rate using discretized features and LDA Classifier

Dataset	Rough Method	Entropy Method	1r method	Chi-merge	Without discretization
Iris	0.0833	0.0540	0.0373	0.0400	0.0200
Glass	0.4317	0.2177*	0.3630	0.3485	0.4149
Diabetes	0.2272	0.2899*	0.2385	0.2282	0.2273
Heartc	0.1606	0.1659*	0.1626	0.1622	0.1643
Ionosfera	0.1341	0.1404	0.1310	0.1601	0.1461
Crx	0.1347	0.1347	0.1356	0.1349	0.1349
Vehicle	0.2823	0.2970	0.2841	0.2911	0.2219
German	0.2432	0.0416	0.2337	0.2414	0.2422

Table III. Subsets of features selected using Rough sets criterion along with five discretization methods

Dataset	Rough Method	1R method	Equal width	ChiMerge	SFS - LDA	SFS – KNN
Glass	4 6 1 2	3 6 1	4 6 1 3	4 6 1 3	4 3 6 2	4 6 3 2
Bupa	4 1 5 3 6	6 1 2 3	4 5 3 6 1	2 5 4 3	3 4 5 6	1 3 5
Heartc	1 5 8 10 4	5,1,10	5 8 4 10 1	5 8 4 1 10	3 9 12 13	2 12 13
Ionosphere	3 1 5 7 8 12 2 10	3 1 19 7 16 6 12 4 14 28	3 1 8 5 7 12 2 10	27 3 31 4 1 6 21 25 19 13 11 29 7 15 26 23 22 8 9 2 12 5 16	3 6 19	1 3 4 14
Diabetes	6 2 7 5	6,8,7,3,4	6 2 7	7 6 5 4 8	2 6 7	2 6 7
Vehicle	8 12 7 9 11 6 10 1 5 13 14	8 7 12	12 7 8 9 11 6 14	7 12 11 8 9 13 3	1 3 4 5 6 8 10 11 17 18	2 5 6 8 9 10

Table IV. Misclassification error rate for the LDA classifier after feature selection

Dataset	Rough Method	1R method	Equal width	Chi-Merge	SFS	Without Sel.
Glass	0.4373	0.4663	0.4242	0.4205	0.4158	0.4135
Bupa	0.3176	0.4197	0.3173	0.3286	0.3257	0.3182
Heartc	0.2777	0.3131	0.2767	0.2804	0.1569	0.1663
Ionosphere	0.1723	0.1658	0.1720	0.1586	0.1831	0.1433
Diabetes	0.2294	0.3108	0.2300	0.3105	0.2295	0.2273
Vehicle	0.2925	0.5855	0.4085	0.3858	0.2426	0.2202

Table V. Misclassification error rate for the KNN classifier after feature selection

Dataset	Rough Method	1R method	Equal width	Chi-Merge	SFS	Without Sel.
Glass	0.3228	0.3771	0.3509	0.3588	0.3158	0.3294
Bupa	0.3521	0.4657	0.3579	0.3562	0.3455	0.3402
Heartc	0.3569	0.4390	0.3558	0.3609	0.1831	0.3474
Ionosphere	0.1301	0.1390	0.1324	0.1384	0.0806	0.1547
Diabetes	0.2845	0.3575	0.2722	0.3208	0.2714	0.2859
Vehicle	0.3823	0.4601	0.4111	0.4248	0.2971	0.3503

Table VI. Misclassification error rates for the KNN classifier using the selected cases

	Rough Set		Chi Merge	
Dataset	Before	After	Before	After
Iris	0.0466	0.08	0.04	0.0488
Heartc (1 4 5 8 10)	0.3692	0.3696	0.3662	0.3741
Ionosfera	0.1542	0.1514	0.1476	0.1714
Crx (2,3,6,8,14,15,16)	0.2215	0.2517	0.3169	0.3230
Diabetes	0.3043	0.2952	0.2834	0.2673
Vehicle	0.3608	0.3940	0.3513	0.3901
Glass	0.3656	0.4562	0.3515	0.3937
German	0.3490	0.3443	0.3420	0.3430

VII. Conclusion

The results obtained on the previous section lead us to the following conclusions.

- Rough set is a good option to data preprocessing tasks in the KDD process.
- Discretization based on Rough sets theory compares well with other discretization methods.
- Feature Selection using Rough sets theory is a way to identify relevant features. Only features having a large dependency with the decisional attribute are considered relevant.
- Instance selection using Rough sets concepts shows good results. This is validated by the improvement on the performance of the KNN classifier.

References

[1] Acuña, E. A comparison of filters and wrappers for feature selection in supervised classification. Proceedings of the Interface 2003 Computing Science and Statistics. 2003, Vol 34.

[2] Acuña, E., and Rodriguez, C. Dprep: Data preprocessing and visualization functions for classification. R package 1.0. http://math.uprm.edu/~edgar/dprep.html. accessed on February 17, 2006.

[3] Cano, J., Herrera F., and Lozano, M. Using evolutionary algorithm as instance selection for data reduction in KDD: An experimental study. In

IEEE transactions on evolutionary computation. 2003, Vol. 7. No. 6. pp. 561-575.

[4] Dougherty, J. Kohavi, R. and Sahami, M. et al. Supervised and Unsupervised Discretization of Continuous Features. Proceeding of twelfth International Conference, Morgan Kaufmann Publishers. 1995, pp. 194 – 202.

[5] Duntsch, I. and Gediga, G. Rough set data analysis. In Encyclopedia of Computer Science and Technology, Marcel Dekker. 2000, 282-301.

[6] Grzymala, J. and Ziarko, W. Data mining and rough set theory. Communications of the ACM. 2000.

[7] Grzymala, J. and Siddhave, S. Rough set Approach to Rule Induction from Incomplete Data. Proceeding of the IPMU'2004, the10th International Conference on information Processing and Management of Uncertainty in Knowledge-Based System. 2004.

[8] Guyon, I, and Elisseeff, A. An introduction to variable and feature selection. Journal of Machine Learning Research. 2003, pp 1157-1182.

[9] Hall, M. Feature Selection for Discrete and Numeric Class Machine Learning. Proc. Seventeenth International conference on Machine Learning, San Francisco, CA, Morgan. Kaufmann. 2000, pp. 359-366.

[10] John, G. Irrelevant Feature and the subset selection problem. In machine Learning: Proceeding of the Eleventh International Conference. In Morgan Kaufmann Publisher. 1994, pp. 121-129.

[11] Kohavi, R. A third dimension to rough sets. In proceeding of the Third international workshop on rough sets and soft computing. 1994, pp. 244-251.

[12] Kohavi, R. and Frasca, B. Useful Feature Subsets and Rough Set Reducts. In proceeding of the Third International Workshop on rough sets and soft computing. San Jose, California. 1994, pp. 310-317.

[13] Kusiak, A. Rough Set Theory: A Data Mining Tool for Semiconductor Manufacturing. IEEE Transactions on electronics Packaging Manufacturing. 2001.

[14] Lesh N., Zaki M, and Ogihara M., "Mining Features for Sequence Classification," 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Diego, CA, August 1999.

[15] Liu, H., Hussain, F., Lim, C. and Dash, M. Discretization: An Enabling Technique. Data Mining and Knowledge Discovery. 2002, pp. 393 - 423.

[16] Liu, H. and Motoda, H. On issues of instance selection. In Data Mining and Knowledge Discovery. 2002, pp.115-130.

[17] Nguyen, S., Nguyen, T., Skowron A. and P. Synak, Knowledge discovery by rough set methods. In: Nagib C. Callaos (eds.), ISAS-96: Proc. of the International Conference on Information Systems Analysis and Synthesis. 1996, pp. 26-33.

[18] Nguyen, S., Skowron A., Synak P., Wróblewski J., Knowledge Discovery in Databases: Rough Set Approach., Proceedings of the Seventh International Fuzzy Systems Association World Congress (IFSA'97). 1997, pp. 204 - 209.

[19] Nguyen, S., Skowron, A. and Stepaniuk, J., Granular Computing: a rough set approach, Computational Intelligence. 2001, pp.514-544.

[20] Ohrn, A. Discernibility and Rough Sets in Medicine: Tools and applications. PhD thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway. 1999.

[21] Pawlak, Z., Grzymala, J., Slowinski, R., and Ziarko, W., Rough Sets. Communications of the ACM. 1995, pp 88-95.

[22] Pawlak, Z., Rough sets and Fuzzy sets. Proceedings of the 1995 ACM 23rd annual conference on computer science. 1995, pp 252-254.

[23] Pudil, P., Ferri, F.J., Novovicova, J., Kittler, J. Floating search methods for feature selection with monotonic criterion function. International Conference on Pattern Recognition. 1994, pp. 279 - 283.

[24] Shen, Q. and Chouchoulas, A. Rough set-based dimensionality reduction for supervised and unsupervised learning. Int. J. Applied Mathematics computational Science. 2001, Vol. 11, No. 3 pp. 583-601.

[25] Shi, H. and Fu, J. A global discretization method based on rough sets. Proceeding of the fourth International Conference on Machine Learning and Cybernetics, Guangzhou. 2005, pp 18-21.

[26] Stepaniuk, J. Rough sets, discretization of attributes and stock market data. Fourth European Congress on Intelligent Techniques and Soft Computing, Proceedings EUFIT'96 , Aachen, Germany, Verlag Mainz. 1996, pp.202-203.

[27] Wang, X., Yang J. Teng X., Xiang, W., Jensen, R., Feature selection based on Rough Sets and particle swarm optimization. Pattern recognition Letters 28.2007, pp. 459-471.

[28] Yu, K., Xu, X., Tao, J., Ester, M., and Kriegel, P., Instance selection techniques for memory-based collaborative filtering. Proc. Second SIAM International conference on Data mining. http://www.siam.org/meetings/sdm02/proceedings/sdm02-04.pdf 2002.

[29] Zhong, N. Using Rough Sets with Heuristics for Feature Selection. Journal of Intelligent Information Systems. 2001, pp. 199 - 214.