# Appendix: Rough Hypercuboid Approach for Feature Selection in Approximation Spaces

Pradipta Maji

## 1 Data Sets Used

This section reports some benchmark data sets that are used to evaluate the performance of different methods. While Iris, Satimage, Segmentation, Isolet, Ionosphere, and Multiple Features data sets are downloaded from the *UCI Machine Learning Repository* [1], Breast Cancer, Leukemia, Lung Cancer, Prostate Cancer, and DLBCLNIH data sets are available at the *Kent Ridge Biomedical Data Set Repository* [2].

1) *Iris:* This is a four-dimensional data set containing 50 samples each of three types of Iris flowers. One of the three clusters (class 1) is well separated from the other two, while classes 2 and 3 have some overlap.
2) *Satimage:* The database is a tiny sub-area of a scene, consisting of $82 \times 100$ pixels, each pixel covering an area on the ground of approximately $80 \times 80$ meters. The information given for each pixel consists of the class value and the intensities in four spectral bands, from the green, red, and infra-red regions of the spectrum. The data set contains 6435 examples: 4435 training and 2000 testing, with 36 real valued attributes and 6 classes.
3) *Lung Cancer:* This data set contains 181 tissue samples: 32 training and 149 testing. Among them 31 are malignant pleural mesothelioma and rest 150 adenocarcinoma of the lung. Each sample is described by the expression levels of 12533 genes.
4) *Breast Cancer:* In this data set, relapse or non relapse of metastases in patients after initial diagnosis for interval of at least 5 years has been classified in breast cancer patients. Total 97 samples are given: 78 training and 19 testing, with 46 patients developed distance metastases within 5 years, labeled as relapse, while 51 remained healthy, labeled as non-relapse. The data set consists of 24481 genes.
5) *Ionosphere:* It represents autocorrelation functions of radar measurements. The task is to classify them into two classes denoting passage or obstruction in ionosphere. There are 351 instances with 34 continuous valued attributes.
6) *Leukemia:* This data set consists of gene expression profiles of 215 training and 112 testing samples classified into 7 classes, six subtypes of pediatric acute lymphoblastic leukemia and one that contains diagnostic samples that did not fit into any one of the six groups. The data set contains total 12558 genes.
7) *Isolet:* The data set consists of several spectral coefficients of utterances of English alphabets by 150 subjects. There are 617 real valued features with 7797 instances and 26 classes.
8) *DLBCLNIH:* Biopsy samples of diffuse large-B-cell lymphoma from 240 patients were examined for gene expression with the use of DNA microarray and analyzed for genomic abnormalities. The 240 samples were divided into two groups: a preliminary group (training) of 160 patients and a validation group (testing) of 80 patients. Number of microarray features is 7399.
9) *Multiple Features:* Multiple features data set consists of features of handwritten numerals ('0'-'9') extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2,000 patterns) have been digitized in binary images. Total 649 attributes are there in the data set.
10) *Segmentation:* This data set contains instances that are drawn randomly from a database of 7 outdoor images. The images are hand segmented to create a classification for every pixel, where each instance is a $3 \times 3$ region. The data set contains 3310 examples: 210 training and 2100 testing, with 18 continuous attributes and 7 classes.
11) *Prostate Cancer:* This data set contains gene expression profiles of 102 training and 34 testing samples classified into 2 classes, tumor and normal, with 12600 genes. The training set contains 52 prostate tumor samples and 50 normal samples, while there are 25 tumor and 9 normal samples in test data set.

## 2 Algorithms Compared

The paper compares the performance of proposed algorithm with that of various feature selection algorithms such as mutual information based approaches: InfoGain [3] and mRMR framework [4]; rough set based approaches: quick reduct (RSQR) [5], discernibility matrix

• *The author is with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: pmaji@isical.ac.in.*

using genetic algorithm (GADM) [6], [7] and MRMS framework (RSMRMS) [8]; fuzzy-rough set based approaches: quick reduct (FRQR) [9] and mRMR framework (FRmRMR) [10]; and margin based approaches: relevance in estimating features (RELIEF) [11], iterative search margin based algorithm (SIMBA) [12] and the method due to Chen and Wasikowski (FAST) [13].

The proposed algorithm, InfoGain [3], RSQR [5], RSMRMS [8], FRQR [9], FRmRMR [10], RELIEF [11], SIMBA [12], and FAST [13] are implemented in C language, while the source code of mRMR [4] algorithm written in C language is downloaded from http://penglab.janelia.org/proj/mRMR. The source codes of FRmRMR [10] and RSMRMS [8] are available at www.isical.ac.in/∼pmaji/important.html. The results obtained using C codes of RELIEF [11] and SIMBA [12] algorithms are validated by the Matlab toolbox available at https://skydrive.live.com/?cid=843f36fc50a5ece8&id=843F36FC50A5ECE8!124. The rough set exploration system (RSES), available at http://logic.mimuw.edu.pl/∼rses/, is used to select features based on the GADM [6], [7]. All the algorithms are run in Ubuntu 10.04 having machine configuration Pentium D, 2.66 GHz, 2 MB cache, and 4 GB RAM

## 3 Optimum Feature Subset

All the results are presented and compared for optimal feature subsets of different algorithms. Table 5 compares the performance of different rough set models, namely, Pawlak's or classical rough sets [14], neighborhood rough sets [15], fuzzy-rough sets [16], and rough hypercuboid approach, using the proposed Max Relevance-Max Dependency-Max Significance (MRMDMS) criterion based feature selection algorithm. Hence, the optimal feature subset for each rough set model reported in Table 5 is determined based on the stopping criterion (Step 4) of the proposed MRMDMS based feature selection algorithm. Similarly, several existing algorithms such as rough set based quick reduct (RSQR) [5], discernibility matrix using genetic algorithm (GADM) [6], [7], and fuzzy-rough quick reduct (FRQR) [9] have their own stopping criteria to select optimal feature subsets. Hence, the results of these algorithms are also presented for their optimal feature subsets in Tables 6 and 7.

On the other hand, some existing feature selection algorithms compared in Tables 6 and 7, namely, mutual information based InfoGain [3] and Min Redundancy-Max Relevance (mRMR) framework [4], rough set based Max Relevance-Max Significance framework (RSMRMS) [8], fuzzy-rough set based mRMR framework (FRmRMR) [10], and margin based approaches such as relevance in estimating features (RELIEF) [11], iterative search margin based algorithm (SIMBA) [12] and ROC-based feature selection metric (FAST) [13], do not have any specific stopping criterion. Hence, the classification accuracy of support vector machine (SVM) on training data set is considered in these cases to select optimal feature subsets. First, all the features of a given data set are ranked using these algorithms, and then the classification accuracy of the SVM is calculated on training data set for different number of features. Finally, the first "d" features are considered for optimal feature subset for which the SVM attains its highest classification accuracy.

## References


[1] A. Frank and A. Asuncion, "UCI Machine Learning Repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[2] "Kent Ridge Bio-medical Data Set Repository." [Online]. Available: http://sdmc.lit.org.sg/GEDatasets/Datasets.html

[3] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[4] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[5] A. Chouchoulas and Q. Shen, "Rough Set-Aided Keyword Reduction for Text Categorisation," *Applied Artificial Intelligence*, vol. 15, no. 9, pp. 843–873, 2001.

[6] A. Skowron and C. Rauszer, "The Discernibility Matrices and Functions in Information Systems," in *Intelligent Decision Support*, R. Slowinski, Ed. Dordrecht: Kluwer Academic Publishers, 1992, pp. 331–362.

[7] R. W. Swiniarski and A. Skowron, "Rough Set Methods in Feature Selection and Recognition," *Pattern Recognition Letters*, vol. 24, pp. 833–849, 2003.

[8] P. Maji and S. Paul, "Rough Set Based Maximum Relevance-Maximum Significance Criterion and Gene Selection from Microarray Data," *International Journal of Approximate Reasoning*, vol. 52, no. 3, pp. 408–426, 2011.

[9] R. Jensen and Q. Shen, "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 12, pp. 1457–1471, 2004.

[10] P. Maji and S. K. Pal, "Feature Selection Using $f$-Information Measures in Fuzzy Approximation Spaces," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 854–867, 2010.

[11] K. Kira and L. A. Rendell, "The Feature Selection Problem: Traditional Methods and A New Algorithm," in *Proceedings of the 10th National Conference on Artificial Intelligence*. MIT Press, 1992, pp. 129–134.

[12] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin Based Feature Selection: Theory and Algorithms," in *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[13] X.-W. Chen and M. Wasikowski, "FAST: A ROC-Based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[14] Z. Pawlak, *Rough Sets: Theoretical Aspects of Resoning About Data*. Dordrecht, The Netherlands: Kluwer, 1991.

[15] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood Rough Set Based Heterogeneous Feature Subset Selection," *Information Sciences*, vol. 178, pp. 3577–3594, 2008.

[16] D. Dubois and H. Prade, "Rough Fuzzy Sets and Fuzzy Rough Sets," *International Journal of General Systems*, vol. 17, pp. 191–209, 1990.