FUZZY ROUGH SET ASSISTED TECHNIQUES FOR DATA REDUCTION TO ENHANCE PREDICTION PERFORMANCES

The thesis submitted in partial fulfilment

for the Award of Degree

DOCTOR OF PHILOSOPHY

by

PANKHURI JAIN

DEPARTMENT OF MATHEMATICAL SCIENCES

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY)

VARANASI -221005

Roll No: 18121023 **June 2022**

Chapter 8

Conclusion

The entire thesis is summarised in this chapter. The main focus of the research is on feature selection. Rough set theory is one such tool that can be utilized for feature selection without the need for any threshold or user specified parameters. The applicability of rough set theory is limited to crisp valued data. Discretization of real valued datasets in such scenario leads to information loss. Fuzzy rough and intuitionistic fuzzy rough set theories provide data reduction of real valued dataset handling vagueness and uncertainty arisinig in the datasets. The target of this research is on reducing data dimensionality by employing models based on fuzzy rough and intuitionistic fuzzy rough sets. These theories effectively reduce data dimensionality of real valued datasets without requiring any additional parameters.

8.1 Feature Selection

In this, we have thesis discussed the idea of dimensionality reduction utilizing fuzzy rough and intuitionistic fuzzy rough set theories while combating various issues arising on the way. The degree of dependency measure is used to measure subset quality. The divergence based intuitionistic fuzzy rough set model is presented for prediction of anti-bacterial peptides for treatment of tuberculosis. Likewise, influence of noise in traditional lower and upper approximations is eliminated by k-mean based intuitionistic fuzzy rough set model. It has effectively diminished the impact of noisy samples. The effectiveness of the k mean based model is further illustrated by its application in predicting aptamer-protein interacting pairs.

The uncertainty and vagueness are handled in much more efficient way by a fitting based intuitionistic fuzzy rough set model. The addon to this model is that it fitted data well and prevented misclassification of samples to incorrect decision class. Neighbourhood concept is utilized to form parameterized granules which is henceforth used for formulating lower and upper approximations and thereby the dependency. The comparative analysis with existing works has demonstrated the performance achieved by the approach.

An extension of feature selection approaches in incomplete datasets is also discussed in the thesis. A fuzzy rough set based approach is introduced to impute missing values followed by feature selection using search optimization techniques. The entire idea has provided a platform for enhancing classifier or prediction performance. It is further beneficial because of selection of optimal features using monarch butterfly search heuristic. Additionally, experimental analysis has illustrated the success and superiority of the discussed method.

Further, a complete data reduction approach involving both instance and feature selection is introduced. The proposed study effectively reduced the data size as

well as dimensionality simultaneously using particle swarm search heuristic. The promising results obtained from experimentation have very well reflected the success of the underlying model. Further, the model is used to solve complex problem of cancer treatment. Another study has discussed the problem of feature selection in unsupervised domain. Two fuzzy rough set based approaches employing dependency degree and boundary measure are proposed in this study for feature selection in unsupervised domain. The candidate feature subsets are selected using proposed earthworm search strategy. The use of search strategies has further enhanced the performance. The feature subset quality is evaluated considering fuzziness arising in the real world applications.

8.2 Future Work

The entire thesis has concentrated on enhancing model's efficiency and tractibility by constructing or forming reduced representation of dataset using feature selection techniques. There exist lot of future possibilities in this area. Specifically, the use of incremental algorithm for handling the thousands of data flooding the market, servers, etc. Also, incremental algorithm will be helpful in handling streaming features (when all the features are not available at particular span of time and may arrive after sometime repeatedly). The problem of big data or huge dataset may somewhat be reduced by deployment of such algorithms.

Other domains like image recognition, signal processing will also benefit from application of feature selection techniques based on fuzzy set theory. Such application domains will be investigated as part of future work. Handling noise, outliers, etc and other pre-processing techniques will also be discussed.

Summarizing, the impact of feature selection or pre-processing techniques is worth

paying heed particularly because of explosive growth of data. The utility of feature selection based on fuzzy rough set has been illustrated in this research work employing series of experimental and comparative analysis. Other factors like noise, missing values, outliers affecting prediction and also impacting feature selection methodologies have also been discussed.

****