Rough Sets in Hybrid Soft Computing Systems

Renpu Li, Yongsheng Zhao, Fuzeng Zhang, and Lihua Song

School of Computer Science and Technology, Ludong University,
Yantai 264025, China
lip0109@sina.com

Abstract. Soft computing is considered as a good candidate to deal with imprecise and uncertain problems in data mining. In the last decades research on hybrid soft computing systems concentrates on the combination of fuzzy logic, neural networks and genetic algorithms. In this paper a survey of hybrid soft computing systems based on rough sets is provided in the field of data mining. These hybrid systems are summarized according to three different functions of rough sets: preprocessing data, measuring uncertainty and mining knowledge. General observations about rough sets based hybrid systems are presented. Some challenges of existing hybrid systems and directions for future research are also indicated.

1 Introduction

Data mining is a process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases [1]. Soft computing [2], a consortium of methodologies in which fuzzy sets, neural networks, genetic algorithms and rough sets are principle members, has been widely applied to deal with various problems that contain uncertainty or imprecision in many fields, especially in data mining [3].

Each soft computing methodology has its own powerful properties and offer different advantages. For example, Fuzzy sets are famous at modeling human reasoning and provide a natural mechanism for dealing with uncertainty. Neural networks are robust to noise and have a good ability to model highly non-linear relationships. Genetic algorithm is particularly useful for optimal search. Rough sets are very efficient in attribute reduction and rule extraction.

On the other hand, these soft computing techniques also have some restrictions that do not allow their individual application in some cases. Fuzzy sets are dependent on expert knowledge. The training times of neural networks are excessive and tedious when the input data are large and most neural network systems lack explanation facilities. The theoretical basis of genetic algorithm is weak, especially on algorithm convergence. Rough sets are sensitive to noise and have the NP problems on the choice of optimal attribute reduct and optimal rules.

In order to cope with the drawbacks of individual approaches and leverage performance of data mining system, it is natural to develop hybrid systems by integrating two or more soft computing technologies. In the last decades research on hybrid soft computing systems concentrates on the combination of neural networks, fuzzy sets

R. Alhajj et al. (Eds.): ADMA 2007, LNAI 4632, pp. 35–44, 2007.

and genetic algorithms [4], and the most notable achievement is neuro-fuzzy computation [5]. Comparatively, hybrid systems based on rough sets, which has been categorized as a soft computing technology only in the recent years, is scarce.

The rest of this paper is organized as follows: In section 2 rough sets and soft computing are briefly introduced. In section 3 hybrid soft computing systems based on rough sets are summarized according to three different functions of rough sets in these systems: preprocessing data, measuring uncertainty and mining knowledge. Some generalized observations are provided in section 4. Finally challenges of existing hybrid systems based on rough sets and directions for future research are indicated in section 5.

2 Rough Sets and Soft Computing

Rough set theory, which was introduced by Pawlak [6] in the early 1980s, is a mathematical approach that can be employed to handle imprecision, vagueness and uncertainty.

Rough sets have many important advantages for data mining [7], such as providing efficient algorithms for finding hidden patterns in data, finding minimal sets of data, generating sets of decision rules from data, and offering straightforward interpretation of obtained results.

In the last two decades, rough sets have widely been applied to data mining and rapidly established themselves in many real-life applications such as medical diagnosis, control algorithm acquisition and process control and structural engineering.

Soft computing, a term coined by Zadeh [2], is a consortium of methodologies that works synergetically and provides in one form or another flexible information processing capability for handling real life ambiguous situations. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, and low-cost solutions. The guiding principle is to devise methods of computation that lead to an acceptable solution at low cost by seeking for an approximate solution to a problem, possibly formulated in an imprecisely way.

Fuzzy logic, neural networks and Probabilistic reasoning are the initial components of soft computing. Later other methodologies like genetic algorithm, chaotic system, and rough sets became members of soft computing one by one. Traditional research of soft computing concentrates on the combination of neural networks, fuzzy sets and genetic algorithm [4], among them neuro-fuzzy computation is the most notable achievement [5].

It is only in recent years that rough sets are studies as a soft computing technology along with other soft computing technologies. As Pawlak said, "…the theory (rough set theory) is not competitive but complementary to other methods and can also be often used jointly with other approaches (e.g. fuzzy sets, genetic algorithms, statistical methods, neural networks etc.)" [8]. The combination of rough sets and other soft computing technologies such as fuzzy sets, genetic algorithms and neural networks has attracted much attention [9], [10] and is growing into an important issue of soft computing.

3 Hybrid Soft Computing Systems Based on Rough Sets

A data mining process can be viewed as consisting of three phases, shown in Fig. 1. Data preprocessing converts raw data into input data of mining algorithm. Many tasks, such as attribute reduction and data discretization, can be included in this phase. An efficient data preprocessing phase not only provides data that can be handled by mining algorithm, but improves efficiency and performance of data mining system by removing redundant or irrelevant data. Mining algorithm is the most important part in the whole process, which extracting patterns from preprocessed data. Pattern is comprehensible for experts of the data mining system, but not for common users. So the phase of interpretation and visualization are needed. By this phase pattern is transformed into knowledge that can be easily understood and applied by users.

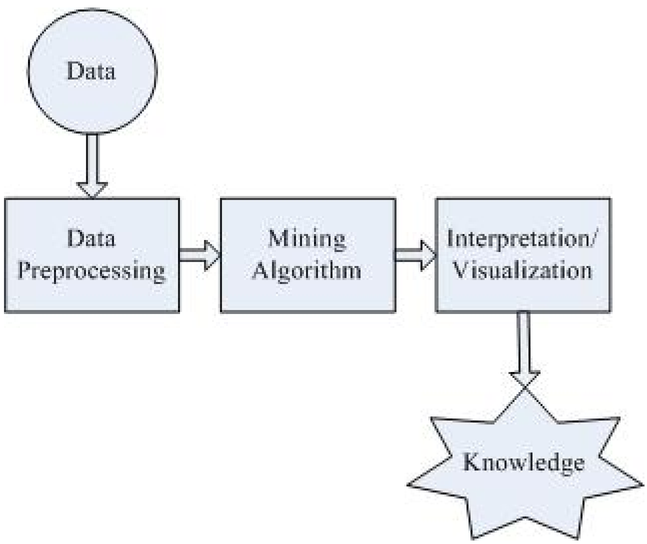

Fig. 1. Overview of a data mining process

Based on the available literatures on rough sets based hybrid soft computing systems, the functions of rough sets in a data mining process can be classified into three categories: preprocessing data, measuring uncertainty, mining knowledge. Each category is described below along with a survey of the existing hybrid systems belong to the category.

3.1 Attribute Reduction by Rough Sets

In this type of hybridization, rough sets are usually used to reduce the size of a dataset by removing some redundant or irrelevant attributes while preserving the classification performance of dataset.

Attribute reduction plays an important role in data mining. Irrelevant and redundant attributes generally affect the performance of mining algorithms. A good choice of a useful attribute subset helps in devising a compact and efficient learning algorithm as well as results in better understanding and interpretation of data in data mining.

Rough sets can be a useful tool for pre-processing data for neural network by attribute reduction to reduce the network's input vector. In [11], a hybrid model integrating rough sets and an artificial neural network is presented to forecast an air-conditioning load. Rough sets are applied to find relevant factors to the load, which are used as inputs of an artificial neural-network to predict the cooling load. Experimental results from a real air-conditioning system show that the hybrid forecasting model is better than comparable models and its relative error is within 4%.

Similarly, in this context rough sets can be integrated with genetic algorithm [12], [13], [14]. In [12], a hybrid approach is proposed to bankruptcy prediction. It firstly uses a rough sets model to identify subsets of potentially important explanatory variables, and then a genetic programming algorithm is applied to construct a bankruptcy prediction model based on these variables. The experimental results show that this hybrid system is more efficient and more effective than an original rough sets model.

Sometimes the process of attribute reduction was accomplished by two steps for addressing noisy data [15], [16]. In the first step rough sets are used to delete irrelevant and redundant attributes, and in the second step neural networks are used to delete noisy attributes.

Two constraints should be considered in a rough set approach for attribute reduction [10]. First, finding a minimal reduct is NP-hard. Second, real-valued attributes can have a very large set of possible values. For the first constraints, Li et al. [17] used genetic algorithm to get the minimal reduct of attributes and Slezak et al. [18] also provided a method of order based genetic algorithm to find optimal approximate entropy reduct. The second constraint was solved in [19] through reducing attribute value field by a rough set based algorithm.

Rough sets can only deal with discrete values. However, real-world data are usually continuous values. Thus it needs a step to transform the continuous valued attributes to discrete ones, namely it should partition the domains of every continuous valued attribute into intervals. In [9] Dependency factor of rough sets is used as the fitness function of GA to get the optimal cutpoints of intervals in order to ensure the maximum consistency of the discrete data.

3.2 Measuring Uncertainty in Data by Rough Sets

In this context rough sets are not the main technique to directly extract knowledge from data, but are used as an assistant tool to estimate some system parameters by its analysis capability for uncertainty in data.

The choice of architecture is a difficult task for neural networks. Rules generated from data by rough sets can be used to determine the architecture of a neural network, including the number of hidden layers, the number of hidden layer nodes and the initial weights of networks [20], [21], [22], [23]. This hybridization helps neural networks possess the characteristics of short learning time, good understandability and strong generalization ability [24], [25].

Qin and Mao [20] proposed an algorithm based on rough sets for fixing the optimal number of hidden layer units. The results on Chinese medicine practical projects show that this algorithm is of wide applicability for the study of neural network architecture.

Yasdi [21] used rough sets for the design of knowledge–based networks in the rough-neuro framework. This methodology consists of generating rules from training examples by using rough sets concepts and mapping them into a single layer of connection weights of a four-layered neural network. In this rough-neuro framework, rough sets were used to speed up or simplify the process of using neural networks. Pal et al. [22] designed a rough-fuzzy MLP for voice recognition. The rules obtained by rough sets are encoded into a neural network, while the confidences of the rules are used to initial the weights of the neural network. The results show that the performance of rough-fuzzy MLP is superior to that of fuzzy MLP and that of MLP without prior knowledge.

On the other hand, rough sets also can be used to adjust the structure of a trained neural network [26], [27]. In [26], the structure of a trained neural network is optimized by rough sets. Dependency of rough set rule is used in an iterative algorithm until a minimal number of rules of the model are selected. Hassan and Tazaki [27] present a hybrid decision support system for medical diagnosis in which rough sets are used to delete the nodes of a neural network.

Fuzzy sets are dependent on expert knowledge, while rough sets can deal with data without any preliminary or additional information. So introduction of rough sets can enlarge the application area of fuzzy sets. Base on rough sets [28] proposes a parameter-free roughness measure for fuzzy sets, which does not depend on parameters that are designed as thresholds of definiteness and possibility in membership of the objects to a fuzzy set.

3.3 Mining Knowledge by Rough Sets

Rule is the most common form of knowledge generated by rough sets based method. Many rule generation algorithms based on rough sets have been proposed to obtain the final knowledge in hybrid soft computing systems [16].

When rough sets is used as the main technique of mining rules from data, other soft computing technologies is also used as the assistant tools to deal with some problems that rough sets can not accomplish.

Some rule extraction algorithms based on rough sets may produce large amounts of rules that are not all useful according to some measures such as accuracy and coverage, and the work picking out the interesting ones from these rules is time consuming. In order to address this problem, Huang and Dai [29] presented a hybrid system based on rough sets and genetic algorithms, rough set is used to extract rules and genetic algorithm is exploited to find the optimal probabilistic rules that have the highest accuracy and coverage, and shortest length. Experimental results show that it run more efficiently than those traditional methods.

In [30], a hybrid scheme based on fuzzy sets and rough sets are proposed for breast cancer detection. Fuzzy sets are firstly used to pre-processing breast cancer images for enhancing the contrast of images. Rough sets-based approaches are applied for attribute reduction and rule extraction. Experimental results show that the hybrid scheme performs well reaching over 98% in overall accuracy. In [31] fuzzy sets are introduced to a rough sets-based information measure for getting the reducts with a better performance.

In order to generating from data with quantitative values which are common in real-world application, Hong et al. [32] proposed an algorithm to produce a set of maximally general fuzzy rules for an approximate coverage of training examples based on the variable rough set model form noisy quantitative training data. Each quantitative value is firstly transformed into a fuzzy set of linguistic terms using membership fuctions. Then fuzzy rules are generated by a rough set-based method.

Rule generation from artificial neural networks has received a great deal of research due to its capability of providing some insight to the user about the symbolic knowledge embedded within the network [33].

In order to improve the classification accuracy of rules, many schemas were proposed. Ahn et al. [34] proposed a hybrid intelligent system by combining rough set approach with neural network for predicting the failure of firms based on the past financial performance data. When a new object is predicted by rule set, it is fed into the neural network if it does not match any of the rules. Thus they use rough sets as a preprocessing tool for neural networks. Though this approach can get high classification accuracy, some knowledge in neural networks is still hidden and not comprehensible for user.

It seems that using fuzzy rules instead of classical decision rules may improve classification results. In [35], a hybrid system combing rough sets and fuzzy sets is provided to improve classification process. Fuzzy sets support approximate reasoning and rough sets are responsible for data analyzing and process of automatic fuzzy rules generation. Drwal and Sikora [36] proposed a rough-fuzzy hybridization scheme for case generation. In this method, fuzzy sets are firstly used to represent a pattern in terms of its membership to linguistic variables. This gives rise to efficient fuzzy granulation of the feature space. Then rough sets are used to obtain fuzzy rules that contain informative and irreducible information both in terms of features and patterns. These fuzzy rules represent different clusters in the granular feature space, which can be mapped to different cases by fuzzy membership functions. Experimental results obtained on three UCI data sets show that this method is superior to traditional methods in terms of classification accuracy, case generation, and retrieval times.

4 A General Observation

Generally, in the two former hybridization types rough sets are used as an assistant tool for mining knowledge from data, and in the latter type as a major element of hybrid systems to directly extracting rule knowledge from data.

Combination of rough sets and neural networks is popular for complementary features of the two technologies. Attribute reduction of rough sets can shorten training time of neural networks, and rules extracted from a trained neural network by rough sets will greatly improve the interpretation of knowledge embedded in trained neural network. On the other hand, neural network is also a good candidate helps rough set deal with noise in the data.

When integrated with rough sets, genetic algorithms are involved in various optimization and search processes, especially for the NP problem of rough set-based algorithms, like optimal reduct and optimal rule set. While Fuzzy sets are suitable for handling the problems related to fuzzy cases.

It is no doubt that hybridization of soft computing technologies can improve performance of a data mining system in such areas as tractablility, low cost, and robustness. According to the specific problems to be solved, hybrid systems can be summarized in Table 1 and Table 2. Table 1 shows typical problems of rough sets that are relievable by other soft computing technologies, while obvious problems of other soft computing technologies with corresponding solution of rough sets are listed in Table 2.

Table 1. Problems of rough sets and corresponding solutions provided by other soft computing techologies

Problem	*Solution*
Sensitivity to noise	Filtrating noise in data by neural networks [15, 16]
NP-hard in attribute reduct and rule extraction	GA-based optimation method [17-19, 29]
Low generalization of rule prediction	Predicted with neural networks classifier or integrated with fuzzy rules [34-36]
Difficulty to deal with fuzzy features of data	Fuzzy sets [30-32]

Table 2. Problems of other soft computing tecnnologies and corresponding solutions provided by rough sets

Problem	*Solution*
Time-consuming of neural network training	Reducing data by rough set-based preprocessing [11]
Lack of explanation of trained neural networks	Rule extraction by rough sets from data or from trained neural networks [33]
Uncertainty on initial architecture and parameters of neural networks	Measuring uncertainty by rough sets [20][21][22][23]
Dependence for domain information of fuzzy sets	Rough sets can analyze data without any preliminary or additional information about data [28]
Time-consuming of convergence of genetic algorithm search	Reducing attributes by rough sets[12][13][14]

5 Conclusions

In this paper we presented an overview of hybrid soft computing systems from a rough set perspective. These systems are discussed and summarized from three categories according to different functions of rough sets in a data mining process: preprocessing data, measuring uncertainty, and mining knowledge. From these systems it can be immediate concluded that rough sets are complementary with other soft computing technologies including neural networks, genetic algorithms and fuzzy sets. Hybridization of rough sets and other soft computing technologies is a promising approach for developing robust, high-efficient, and low-cost data mining systems.

Although hybrid soft computing systems based on rough sets have been successfully applied in many areas, some of the challenges still exist. For example:

- Lack of general theoretical framework and design principle for integration of rough sets and other technologies.
- Efficient mechanism to handle complex data which is dynamic, incomplete, sequential, etc.
- Integration of extensive rough set theory and other soft computing methodologies.
- Quantitative evaluation of system performance

Some aspects of hybrid soft computing systems based on rough sets are as follows:

1) At present most systems are based on classical rough set theory. However, many limitations of classical rough set theory have been widely realized and corresponding extensive theory for rough sets have been proposed. How to integrate efficiently extensive rough set theory with neural networks and genetic algorithms needs to be developed in near future. Some attempts in this direction are described in [37].

2) Facing complex data, efficient hybridization mechanism including management for missing values and update of knowledge is a potential issue [38].

3) Interaction with user is of great significance for successful application of a data mining system. So the related researches such as utilization of expert knowledge and visualization of mined knowledge need to be studied more deeply.

4) Novel combinations are still needed for better utilizing the advantages of rough sets and other technologies and developing high-performance, low-cost and robust data mining systems.

Acknowledgement

This work is supported by the Youth Natural Science Foundation of Ludong University under contract 22480301.

References

1. Chen, M., Han, J., Yu, P.S.: Data Mining: An Overview Form a Database Perspective. IEEE Trans. Knowledge and Data Engineering 8, 866–883 (1996)
2. Zadeh, L.A.: Fuzzy Logic, Neural Networks, and Soft Computing. Commun. ACM 37, 77–84 (1994)
3. Mitra, S., Pal, S.K., Mitra, P.: Data Mining in Soft Computing Framework: A Survey. IEEE Trans. Neural Networks 11, 3–14 (2002)

4. Bonissone, P.P., Chen, Y.-T., Goebel, K., et al.: Hybrid Soft Computing Systems - Industrial and Commercial Applications. Proceedings of the IEEE 9, 1641–1667 (1999)
5. Mitra, S., Hayashi, Y.: Neuro-fuzzy Rule Generation: Survey in Soft Computing Framework. IEEE Trans. Neural Networks 11, 748–768 (2000)
6. Pawlak, Z.: Rough Sets. Inter. J. of Computer and Information Sciences 5, 341–356 (1982)
7. Pawlak, Z.: Rough Sets and Intelligent Data Analysis. Information Sciences 147, 1–12 (2002)
8. Pawlak, Z.: Data Mining – A Rough Set Perspective. In: Zhong, N., Zhou, L. (eds.) Methodologies for Knowledge Discovery and Data Mining. LNCS (LNAI), vol. 1574, pp. 3–12. Springer, Heidelberg (1999)
9. Li, Y., Jiang, J.-P.: The Integrated Methodology of Rough Sets Theory, Fuzzy Logic and Genetic Algorithms for Multisensor Fusion. In: Proceedings of the American Control Conference. IEEE Arlington VA, pp. 25–27. IEEE Computer Society Press, Los Alamitos (2001)
10. James, F.P., Marcin, S.S.: Rough Neurocomputing: A Survey of Basic Models of Neurocomputation. In: Alpigini, J.J., Peters, J.F., Skowron, A., Zhong, N. (eds.) RSCTC 2002. LNCS (LNAI), vol. 2475, pp. 308–315. Springer, Heidelberg (2002)
11. Hou, Z., Lian, Z., et al.: Cooling-load prediction by the combination of rough set theory and an artificial neural-network based on data-fusion technique. Applied Energy 83, 1033–1046 (2006)
12. Thomas, E.M., Terje, L.: Genetic Programming and Rough Sets: A Hybrid Approach to Bankruptcy Classification. European Journal of Operational Research 138, 436–451 (2002)
13. Zhai, L.-Y., Khoo, L.-P., Fok, S.-C.: Feature Extraction Using Rough Sets and Genetic Algorithms – An application for the Simplification of Product Quality Evaluation. Computers & Industrial Engineering 43, 661–676 (2002)
14. Khoo, L.-P., Zhai, L.-Y.: A Prototype Genetic Algorithm-enhanced Rough Set-based Rule Induction System. Computers in Industry 46, 95–106 (2001)
15. Chakraborty, B.: Feature Subset Selection by Neuro-rough Hybridization. In: Ziarko, W., Yao, Y. (eds.) RSCTC 2000. LNCS (LNAI), vol. 2005, pp. 519–526. Springer, Heidelberg (2001)
16. Li, R., Wang, Z.: Mining Classification Rules Using Rough Sets and Neural Networks. European Journal of Operational Research 157, 439–448 (2004)
17. Li, Q.D., Chi, Z.X., Shi, W.B.: Application of Rough Set Theory and Artificial Neural Network for load forecasting. In: Proceedings of the first Int. Conf. on Machine Learning and Cybernetics, pp. 1148–1152. IEEE Computer Society Press, Beijing (2002)
18. Slezak, D., Wroblewski, J.: Order Based Genetic Algorithms for The Search of Approximate Entropy Reduct. In: Wang, G., Liu, Q., Yao, Y., Skowron, A. (eds.) RSFDGrC 2003. LNCS (LNAI), vol. 2639, pp. 308–311. Springer, Heidelberg (2003)
19. Jelonek, J., Krawiec, K., Slowinski, R.: Rough Set Reduction of Attributes and Their Domains for Neural Networks. Computational Intelligence 11, 339–347 (1995)
20. Qin, Z., Mao, Z.: A New Algorithm for Neural Network Architecture Study. In: Proceedings of the 3rd World Congress on Intelligent Control and Automation, pp. 795–799. IEEE Computer Society Press, Hefei (2000)
21. Yasdi, R.: Combining Rough Sets Learning and Neural Learning Method to Deal with Uncertain and Imprecise Information. Neurocomputation 7, 61–84 (1995)
22. Pal, S.K., Mitra, S., Mitra, P.: Rough-fuzzy MLP: Modular Evolution, Rule Generation, and Evaluation. IEEE Trans. on Knowledge and Data Engineering 15, 14–25 (2003)

23. Huang, X.M., Yi, J.K., Zhang, Y.H.: A method of constructing fuzzy neural network based on rough set theory. In: International Conference on Machine Learning and Cybernetics, pp. 1723–1728. IEEE Computer Society Press, Xi'an (2003)
24. Tian, K., Wang, J., et al.: Design of a novel neural networks based on rough sets. In: Chinese Control Conference, pp. 834–838. IEEE Computer Society Press, Harbin (2006)
25. Wu, Z.C.: Research on Remote Sensing Image Classification Using Neural Network Based on Rough Sets. In: International Conference on Info-tech and Info-net, pp. 279–284. IEEE Press & PPTPH, Beijing (2001)
26. Yon, J.H., Yang, S.M., Jeon, H.T.: Structure Optimization of Fuzzy-Neural Network Using Rough Set Theory. In: Proceedings of 1999 IEEE International Conference on Fuzzy Systems, pp. 1666–1670. IEEE Computer Society Press, Seoul (1999)
27. Hassan, Y., Tazaki, E.: Decision Making Using Hybrid Rough Sets and Neural Networks. Int. J. of Neural System 12, 435–446 (2002)
28. Huynh, V.N., Nakamori, Y.: An Approach to Roughness of Fuzzy Sets. In: Proceedings of the 13th IEEE International Conference on Fuzzy Systems, pp. 115–120. IEEE Computer Society Press, Budapest (2004)
29. Hang, X., Dai, H.: An Optimal Strategy for Extracting Probabilistic Rules by Combining Rough Sets and Genetic Algorithm. In: Grieser, G., Tanaka, Y., Yamamoto, A. (eds.) DS 2003. LNCS (LNAI), vol. 2843, pp. 153–165. Springer, Heidelberg (2003)
30. Hassanien, A.: Fuzzy rough sets hybrid scheme for breast cancer detection. Image and Vision Computing 25, 172–183 (2007)
31. Hu, Q., Yu, D., Xie, Z.: Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recognition Letters 27, 414–423 (2006)
32. Hong, T.-P., Wang, T.-T., Chien, B.-C.: Learning Approximate Fuzzy Rules from Training Examples. In: Proceedings of the 2001 IEEE International Conference on Fuzzy Systems, pp. 256–259. IEEE Computer Society Press, Melbourne (2001)
33. Fu, L.: Rule Generation from Neural Networks. IEEE Trans. on Systems, Man and Cybernetics 24, 1114–1124 (1994)
34. Ahn, B.S., Cho, S.S., Kim, C.Y.: The Integrated Methodology of Rough Set Theory and Artificial Neural Network for Business Failure Prediction. Expert Systems with Application 18, 65–74 (2000)
35. Pal, S.K., Mitra, P.: Case Generation Using Rough Sets with Fuzzy Representation. IEEE Trans. on Knowledge and Data Engineering 16, 292–300 (2004)
36. Drwal, G., Sikora, M.: Induction of Fuzzy Decision Rules Based Upon Rough Sets Theory. In: IEEE Conf. on Fuzzy Systems, pp. 1391–1395. IEEE Computer Society Press, Gliwice (2004)
37. Su, C.-T., Hsu, J.-H.: Precision parameter in the variable precision rough sets model: an application. Omega 34, 149–157 (2006)
38. Sikder, I., Gangopadhyay, A.: Managing uncertainty in location services using rough set and evidence theory. Expert Systems with Applications 32, 386–396 (2007)