Computational Theory Perception (CTP), Rough-Fuzzy Uncertainty Analysis and Mining in Bioinformatics and Web Intelligence: A Unified Framework

Sankar K. Pal

Center for Soft Computing Research: A National Facility
Indian Statistical Institute, Kolkata - 700108
sankar@isical.ac.in

Abstract. The concept of computational theory of perceptions (CTP), its characteristics and the relation with fuzzy-granulation (f-granulation) are explained. Role of f-granulation in machine and human intelligence and its modeling through rough-fuzzy integration are discussed. The Significance of rough-fuzzy synergestic integration is highlighted through three examples, namely, rough-fuzzy case generation, rough-fuzzy c-means and rough-fuzzy c-medoids along with the role of fuzzy granular computation. Their superiority, in terms of performance and computation time, is illustrated for the tasks of case generation (mining) in large-scale case-based reasoning systems, segmenting brain MR images, and analyzing protein sequences. Different quantitative measures for rough-fuzzy clustering are explained. The effectiveness of rough sets in constructing an ensemble classifier is also illustrated in a part of the article along with its performance for web service classification. The article includes some of the existing results published elsewhere under different topics related to rough sets and attempts to integrate them with CTP in a unified framework providing a new direction of research.

Keywords: soft computing, fuzzy granulation, rough-fuzzy computing, bioinformatics, MR image segmentation, case based reasoning, data mining, web service classification

1 Introduction

Rough set theory [33] is a popular mathematical framework for granular computing. The focus of rough set theory is on the ambiguity caused by limited discernibility of objects in the domain of discourse. Granules are formed as objects and are drawn together by the limited discernibility among them. A rough set represents a set in terms of lower and upper approximations. The lower approximation contains granules that completely belong in the set and the upper approximation contains granules that partially or completely belong in the set. Rough set-based techniques have been used in the fields of pattern recognition

J.F. Peters and A. Skowron (Eds.) Transactions on Rough Sets XI, LNCS 5946, pp. 106–129, 2010

[25,41], image processing [38], data mining and knowledge discovery [5,31] process from large data sets. Recently rough sets were found to have extensive application in dimensionality reduction [41] and knowledge encoding [2,19] particularly when the uncertainty is due to granularity in the domain of discourse. It is also has been found to be an effective machine learning tool for designing ensemble classifier.

Recently rough-fuzzy computing has drawn the attention of researches in machine learning community. Rough-fuzzy techniques are efficient hybrid techniques based on judicious integration of the principles of rough sets and fuzzy sets. While the membership functions of fuzzy sets enables efficient handling of overlapping classes, the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in class definitions. Since the rough-fuzzy approach has the capability of providing a stronger paradigm for uncertainty handling, it has greater promise in application domains e.g., pattern recognition, image processing, dimensionality reduction, data mining and knowledge discovery, where fuzzy sets and rough sets are being effectively used. Its effectiveness in handling large data sets (both in size and dimension) is also evident because of its "fuzzy granulation" characteristics. Some of the challenges arising out of those posed by massive data and high dimensionality, nonstandard and incomplete data, knowledge discovery using linguistic rules and over-fitting problems can be dealt well using soft computing and rough-fuzzy approaches.

The World Wide Web (WWW) and bioinformatics are the two major forefront research areas where recent data mining finds significant applications. A detailed review explaining the state of the art and the future directions for *web mining* research in soft computing framework is provided by Pal et al. [21]. One may note that web mining, although considered to be an application area of data mining on the stocktickerWWW, demands a separate discipline of research. The reason is that web mining has its own characteristic problems (e.g., page ranking, personalization), because of the typical nature of the data, components involved and tasks to be performed, which cannot be usually handled within the conventional framework of data mining and analysis. Moreover, being an interactive medium, human interface is a key component of most web applications.

Bioinformatics can be viewed as a discipline of using *computational methods to make biological discoveries* [1]. It is an interdisciplinary field mainly involving biology, computer science, mathematics and statistics to analyze biological sequence data, genome content and arrangement, and to predict the function and structure of macromolecules. The ultimate goal is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be derived. With the need to handle large heterogeneous data sets in biology in a robust and computationally efficient manner, soft computing, which provides machinery for handling uncertainty, learning and adaptation with massive parallelism, and powerful search and imprecise reasoning, have recently gained the attention of researchers for their efficient mining.

The significance of some of the soft computing tools for bioinformatics research is reported in different surveys [22,35].

While talking about pattern recognition and decision-making in the 21st century, it will remain incomplete without the mention of the *Computational Theory of Perceptions (CTP)* explained by Zadeh [44,45], which is governed by perception-based computation. Since the boundaries of perceptions (e.g., perception of direction, time, speed, age) are not crisply defined and the attributes it can accept are granules, the concept of rough fuzzy computing seems to have a significant role in modeling the f-granulation (i.e., fuzzy-granules) characteristics of CTP. In the present article, we mention some of the results published elsewhere in the areas of rough-fuzzy approach and fuzzy granular computing with application to tasks like case generation, classification, clustering/segmentation in protein sequence and web data, and integrate them with the concept of f-granulation of CTP in a unified framework; thereby showing greater promise of its research.

The organization of this paper is as follows. Section 2 introduces the basic notions of computational theory of perceptions and f-granulation, while Section 3 presents rough-fuzzy approach to granular computation, in general. Section 4 explains the application of rough-fuzzy granulation in case based reasoning where the problem of case generation is considered. Sections 5 and 6 demonstrate the concept of rough-fuzzy clustering and some of the quantative measures for evaluating the performance of clustering. The problem of segmenting brain MR images is considered as an example. Section 7 demonstrates an application of rough-fuzzy clustering for analyzing protein sequence for determining bio-bases. Section 8 deals with rough set theoretic ensemble classifier with application to web services. Concluding remarks are given in Section 9.

2 Computational Theory of Perceptions and F-Granulation

The *computational theory of perceptions* (CTP) [44,45] is inspired by the remarkable human capability to perform a wide variety of physical and mental tasks, that include recognition tasks without any measurements and any computations. Typical everyday examples of such tasks are parking a car, driving in city traffic, cooking meal, understanding speech, and recognizing similarities. This capability is due to the crucial ability of human brain to manipulate perceptions of time, distance, force, direction, shape, color, taste, number, intent, likelihood, and truth, among others.

Recognition and perception are closely related. In a fundamental way, a recognition process may be viewed as a sequence of decisions. Decisions are based on information. In most realistic settings, decision-relevant information is a mixture of measurements and perceptions; e.g., the car is six year old but looks almost new. An essential difference between measurement and perception is that in general, measurements are crisp, while perceptions are fuzzy. In existing

theories, perceptions are converted into measurements, but such conversions in many cases, are infeasible, unrealistic or counterproductive. An alternative, suggested by the CTP, is to convert perceptions into propositions expressed in a natural language, e.g., it is a warm day, he is very honest, it is very unlikely that there will be a significant increase in the price of oil in the near future.

Perceptions are intrinsically imprecise. More specifically, perceptions are f-granular, that is, both fuzzy and granular, with a granule being a clump of elements of a class that are drawn together by indistinguishability, similarity, proximity or functionality. For example, a perception of height can be described as very tall, tall, middle, short, with very tall, tall, and so on constituting the granules of the variable 'height'. F-granularity of perceptions reflects the finite ability of sensory organs and, ultimately, the brain, to resolve detail and store information. In effect, f-granulation is a human way of achieving data compression. It may be mentioned here that although information granulation in which the granules are crisp, i.e., f-granular, plays key roles in both human and machine intelligence, it fails to reflect the fact that, in much, perhaps most, of human reasoning and concept formation the granules are fuzzy (f-granular) rather than crisp. In this respect, generality increases as the information ranges from singular (age: 22 yrs), c-granular (age: 20-30 yrs) to f-granular (age: "young"). It means CTP has, in principle, higher degree of generality than qualitative reasoning and qualitative process theory in AI [12,40]. The types of problems that fall under the scope of CTP typically include: perception based function modeling, perception based system modeling, perception based time series analysis, solution of perception based equations, and computation with perception based probabilities where perceptions are described as a collection of different linguistic *if-then* rules.

F-granularity of perceptions puts them well beyond the meaning representation capabilities of predicate logic and other available meaning representation methods [44]. In CTP, meaning representation is based on the use of so called constraint-centered semantics, and reasoning with perceptions is carried out by goal-directed propagation of generalized constraints. In this way, the CTP adds to existing theories the capability to operate on and reason with perception-based information.

This capability is already provided, to an extent, by fuzzy logic and, in particular, by the concept of a linguistic variable and the calculus of fuzzy if-then rules. The CTP extends this capability much further and in new directions. In application to pattern recognition and data mining, the CTP opens the door to a much wider and more systematic use of natural languages in the description of patterns, classes, perceptions and methods of recognition, organization, and knowledge discovery. Upgrading a search engine to a question- answering system is another prospective candidate in web mining for CTP application. However, one may note that dealing with perception-based information is more complex and more effort-intensive than dealing with measurement-based information, and this complexity is the price that has to be paid to achieve superiority.

3 Granular Computation and Rough-Fuzzy Approach

Rough set theory [33] provides an effective means for analysis of data by synthesizing or constructing approximations (upper and lower) of set concepts from the acquired data. The key notions here are those of "information granule" and "reducts". Information granule formalizes the concept of finite precision representation of objects in real life situation, and reducts represent the core of an information system (both in terms of objects and features) in a granular universe. *Granular computing* (GrC) refers to that domain where computation and operations are performed on information granules (clump of similar objects or points). Therefore, it leads to have both data compression and gain in computation time, and finds wide applications [29]. An important use of rough set theory and granular computing in data mining has been in generating logical rules for classification and association [33]. These logical rules correspond to different important regions of a feature space, which represent data clusters roughly. For example, given the object region in Figure 1, rough set theory can, whether supervised or unsupervised, extract the rule $F_{1M} \wedge F_{2M}$ (i.e., feature F_1 is M AND feature F_2 is M) to encode the object region. This rule, which represents the rectangle (shown by bold line), provides a crude description of the object or region.

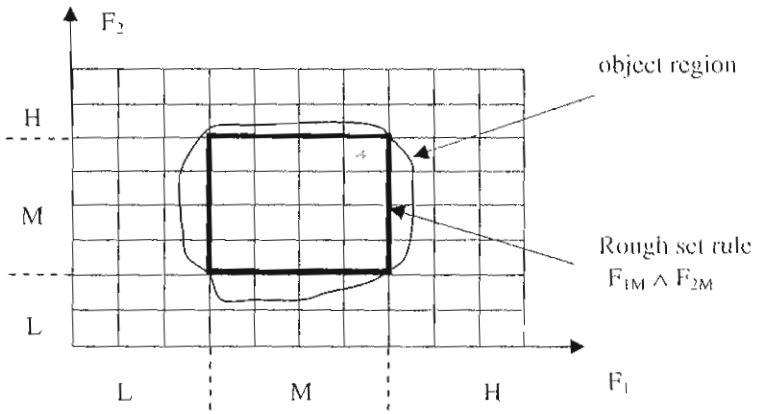

Fig. 1. Rough set theoretic rules for an object

In many situations, when a problem involves incomplete, uncertain and vague information, it may be difficult to differentiate distinct elements and one is forced to consider granules. On the other hand, in some situations though detailed information is available, it may be sufficient to use granules in order to have an efficient and practical solution. Granulation is an important step in the human cognition process. From a more practical point of view, the simplicity derived from granular computing is useful for designing scalable data mining algorithms.

There are two aspects of granular computing: one deals with formation, representation and interpretation of granules (algorithmic aspect) while the other deals with utilization of granules for problem solving (semantic aspect). Several approaches for granular computing have been suggested in literature including fuzzy set theory, rough set theory, power algebras and interval analysis [43,47]. The rough set theoretic approach is based on the principles of set approximation and provides an attractive framework for knowledge encoding and discovery.

For the past few years, rough set theory and granular computation has proven to be another soft computing tool which, in various synergistic combinations with fuzzy logic, artificial neural networks and genetic algorithms, provides a stronger framework to achieve tractability, robustness, low cost solution and close resembles with human like decision making [27,29,31,46]. For example, rough-fuzzy integration can be considered as a way of emulating the basis for f-granulation in CTP, where perceptions have fuzzy boundaries and granular attribute values. Similarly, rough neural synergistic integration helps in extracting crude domain knowledge in the form of rules for describing different concepts/classes, and then encoding them as network parameters; thereby constituting the initial knowledge-base network for efficient learning. Since, in granular computing, computations/operations are performed on granules (clump of similar objects or points) rather than on the individual data points, the computation time is greatly reduced. The results on these investigations, both theory and real life applications, are being available in different journals and conference proceedings [32,42]. Some special issues and edited volumes have also come out [23,24,25]. Rough-fuzzy computing is one of the hybridization techniques that has drawn the attention of researcher in recent times as they promise to provide a much more stronger paradigm for uncertainty handling than the individuals ones. Recently a generalized rough set is defined by Sen and Pal [39] for uncertainty handling and defining rough entropy based on the four criteria, namely, (i) set is crisp and granules are crisp, (ii) set is fuzzy and granules are crisp, (iii) set is crisp and granules are fuzzy, and (iv) set is fuzzy and granules are fuzzy. The f-granulation property of CTP can therefore be modeled using the rough-fuzzy computing framework with one or more of the aforesaid criteria.

Two examples of rough fuzzy computing in case-based reasoning and clustering are explained in the following two sections together with their characteristic features. In the former case granules are fuzzy and the classes are crisp, while the cases are fuzzy and granules are crisp in the latter case. Application of rough-fuzzy clustering in bioinformatics is mentioned in Section 7, as an example of amino acid sequence analysis for determining bio-bases.

4 Rough-Fuzzy Granulation and Case Based Reasoning

Case-based reasoning (CBR) [10], which is a novel Artificial Intelligence (AI) problem-solving paradigm, involves adaptation of old solutions to meet new

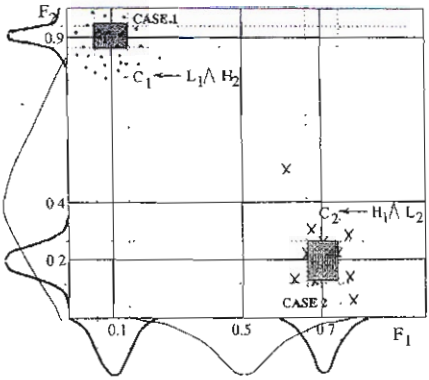

Fig. 2. Rough-fuzzy case generation for a two dimensional data

demands, explanation of new situations using old instances (called cases), and performance of reasoning from precedence to interpret new problems. It has a significant role to play in today's pattern recognition and data mining applications involving CTP, particularly when the evidence is sparse. The significance of soft computing to CBR problems has been adequately explained in a recent book by Pal, Dillon and Yeung [26] and Pal and Shiu [27]. In this section we give an example [28,29] of using the concept of f-granulation, through rough-fuzzy computing, for performing an important task, namely, *case generation*, in large scale CBR systems.

A case may be defined as a contextualized piece of knowledge representing evidence that teaches a lesson fundamental to achieving goals of a system. While case selection deals with selecting informative prototypes from the data, case generation concerns the construction of 'cases' that need not necessarily include any of the given data points. For generating cases, linguistic representation of patterns is used to obtain a fuzzy granulation of the feature space. Rough set theory is used to generate dependency rules corresponding to informative regions in the granulated feature space. The fuzzy membership functions corresponding to the informative regions are stored as cases. Figure 2 shows an example of such case generation for a two dimensional data having two classes. The granulated feature space has $3^2 = 9$ granules. These granules of different sizes are characterized by three membership functions along each axis, and have ill-defined (overlapping) boundaries. Two dependency rules: $class_1 \leftarrow L_1 \wedge H_2$ and $class_2 \leftarrow H_1 \wedge L_2$ are obtained using rough set theory. The fuzzy membership functions, marked bold, corresponding to the attributes appearing in the rules for a class are stored as its case.

Unlike the conventional case selection methods, the cases, illustrated in Figure 2, are cluster granules and not sample points. Also, since all the original features may not be required to express the dependency rules, each case involves a reduced number of relevant features. The methodology is therefore suitable

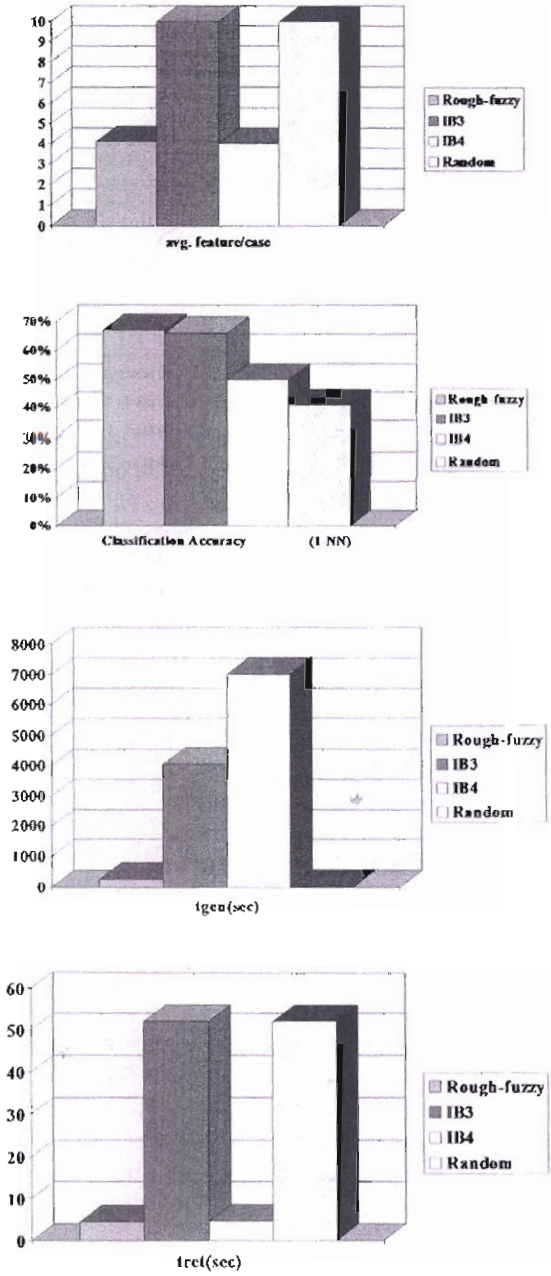

Fig. 3. Performance of different case generation schemes for the forest cover-type GIS data set with 7 classes, 10 features and 586012 samples

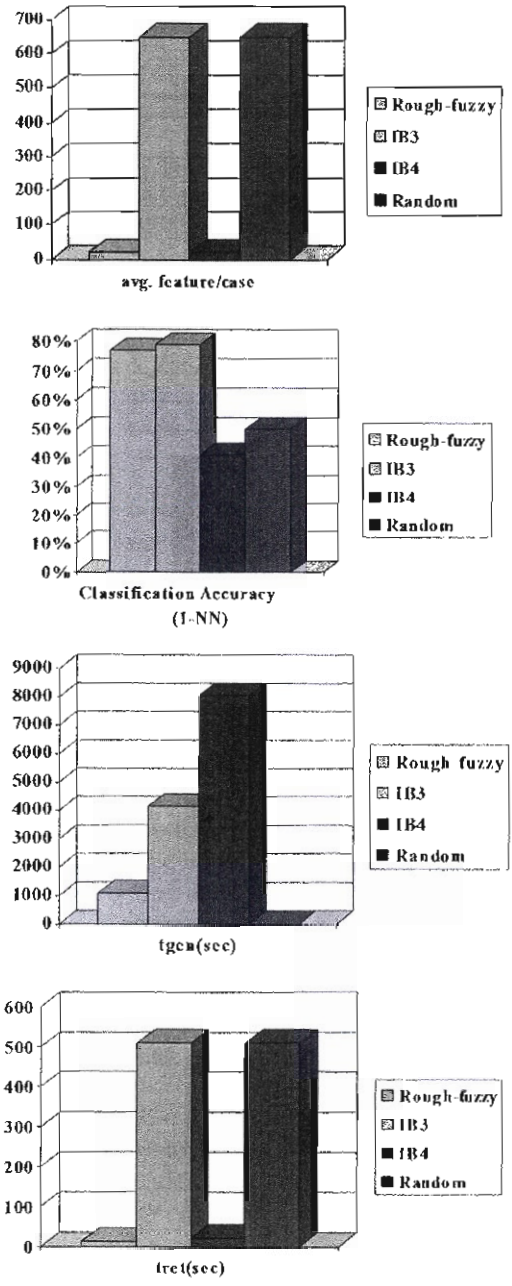

Fig. 4. Performance of different case generation schemes for the handwritten numeral recognition data set with 10 classes, 649 features and 2000 samples

for mining data sets, large both in dimension and size, due to its low time requirement in case generation as well as retrieval.

The aforesaid characteristics are demonstrated in Figures 4 and 4 [28,29] for two real life data sets with features 10 and 649 and number of samples 586012 and 2000, respectively. Their superiority over IB3, IB4 [10] and random case selection algorithms, in terms of classification accuracy (with one nearest neighbor rule), case generation (t_{gen}) and retrieval (t_{ret}) times, and average storage requirement (average feature) per case, are evident. The numbers of cases considered for comparison are 545 and 50, respectively. Recently, Li et al reported a CBR-based classification system combining efficient feature reduction and case selection based on the concept of rough sets [13].

5 Rough-Fuzzy Clustering

Incorporating both fuzzy and rough sets, a new clustering algorithm is described here. This method adds the concept of fuzzy membership of fuzzy sets, and lower and upper approximations of rough sets into a clustering algorithm that results in c number of clusters. While the membership of fuzzy sets enables efficient handling of overlapping partitions, rough sets deal with uncertainty, vagueness, and incompleteness in class definition [15]. In other words, fuzziness is involved here not in determining granules (unlike the case-based method of Section 4), but in handling uncertainty arising from overlapping regions.

Here each cluster is represented by a centroid, a crisp lower approximation, and a fuzzy boundary. The lower approximation influences the fuzziness of a final partition. According to the definitions of lower approximations and boundary of rough sets, if an object belongs to lower approximations of a cluster, then the object does not belong to any other clusters. That is, the object is contained in that cluster definitely. Thus, the weights of the objects in lower approximation of a cluster should be independent of other centroids and clusters, and should not

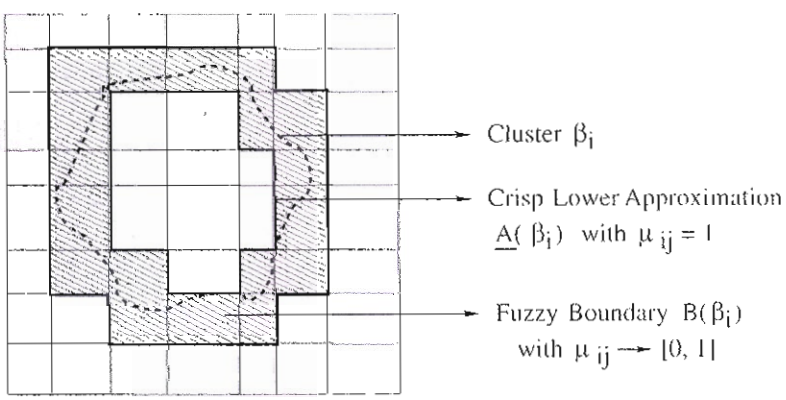

Fig. 5. Rough-fuzzy c-means: each cluster is represented by crisp lower approximations and fuzzy boundary

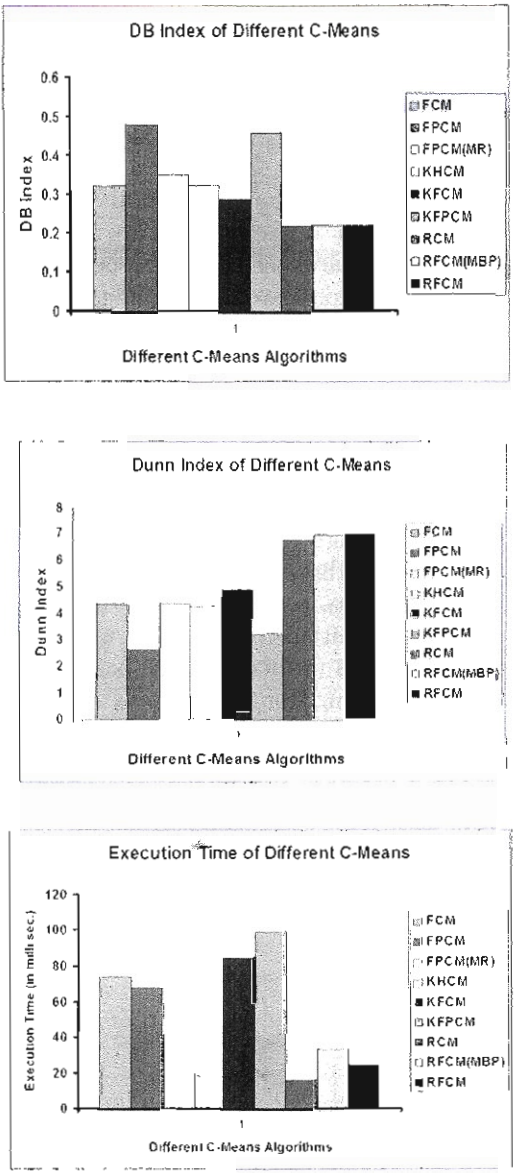

Fig. 6. Comparison of DB and Dunn Index, and execution time of HCM, FCM [3], RCM [14], RFCMMBP [18], and RFCM for Iris Data

be coupled with their similarity with respect to other centroids. Also, the objects in lower approximation of a cluster should have similar influence on the corresponding centroids and cluster. Whereas, if the object belongs to the boundary of a cluster, then the object possibly belongs to that cluster and potentially

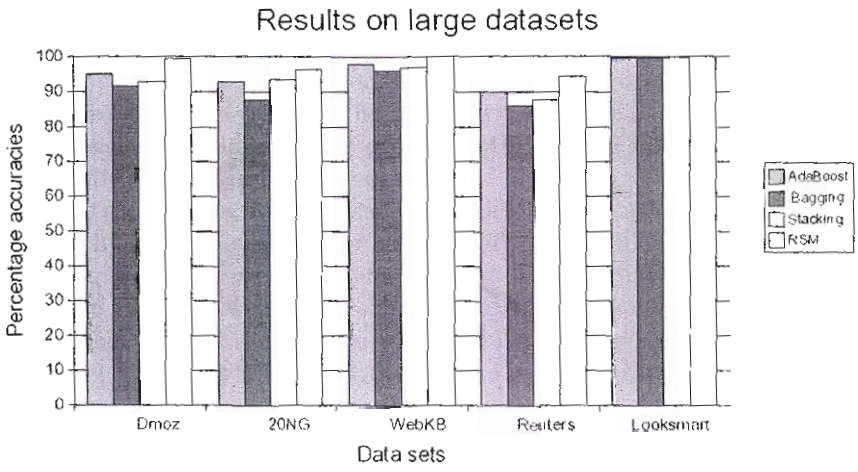

Fig. 7. Comparison of β index of HCM, FCM [3], RCM [14], RFCMMBP [18], and RFCM

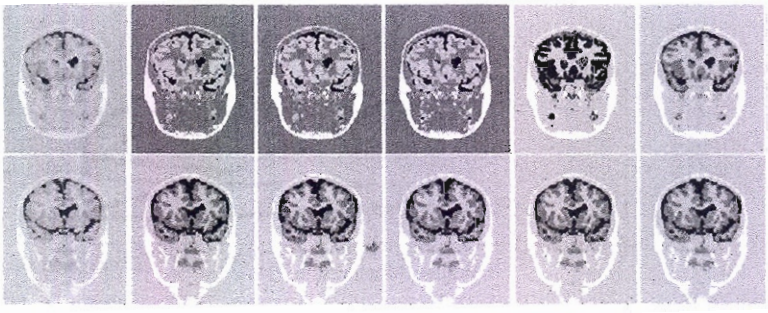

Fig. 8. Some original and segmented images of HCM, FCM [3], RCM [14], RFCMMBP [18], and RFCM

belongs to another cluster. Hence, the objects in boundary regions should have different influences on the centroids and clusters.

So, in the case of rough-fuzzy c-means algorithm (RFCM), the membership values of objects in lower approximation are 1, while those in boundary region are the same as fuzzy c-means. In other word, RFCM first partitions the data into two classes - lower approximation and boundary. Only the objects in boundary are fuzzified. The new centroid is calculated based on the weighting average of the crisp lower approximation and fuzzy boundary. Computation of the centroid is modified to include the effects of both fuzzy memberships and lower and upper bounds. *In essence, Rough-Fuzzy clustering tends to compromise between restrictive (hard clustering) and descriptive (fuzzy clustering) partitions.*

The effectiveness of the algorithm is shown, as an example, for classification of Iris data set and segmentation of brain MR images where the centroid of a

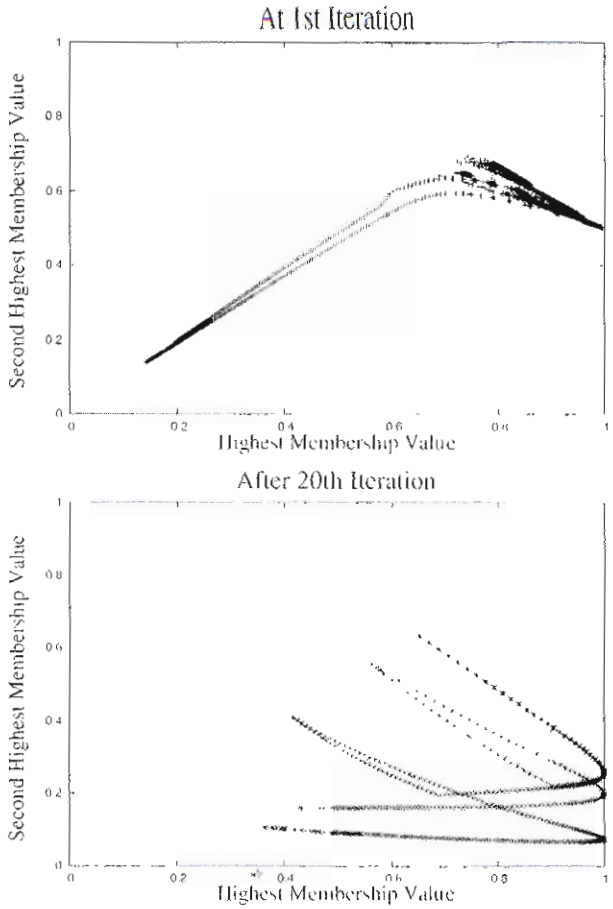

Fig. 9. Scatter plots of two highest membership values of all the objects in the data set of image

cluster is meant for cluster mean, i.e., rough-fuzzy c-means (RFCM). The Iris data set is a four-dimensional data set containing 50 samples of 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.

The performance of other different c-means algorithms is shown with respect to DB and Dunn index [3] in Fig. 4. The results reported establish the fact that RFCM provides best result having lowest DB index and highest Dunn index with lower execution time.

For segmentation of brain MR images, 100 MR images with different sizes and 16 bit gray levels are tested. All the MR images are collected from Advanced Medicare and Research Institute (AMRI), Kolkata, India. The comparative performance of different c-means is shown in Fig. 7 with respect to β index [30]. β index is defined as the ratio of total variation to cluster variation of intensity in

an image. Therefore, for a given number of clusters in an image, higher β-value is desirable.]

Some of the original images along with their segmented versions with different c-means are shown in Fig. 8. The results explain that the RFCM algorithm produces segmented images more promising than do the conventional methods, both visually and in terms of β index.

Figure 9 shows the scatter plots of the highest and second highest membership of all the objects in the data set of image at first and final iterations respectively, considering w=0.95, ($m_1 = 2.0,$) and $c = 4$. The diagonal line represents the zone where two highest memberships of objects are equal. From Fig. 9, it is observed that though the average difference between two highest memberships of the objects are very low at first iteration $\delta = 0.145$), they become ultimately very high at the final iteration $\delta = 0.652$).

6 Quantitative Measures

In this section we describe some quantitative indices to evaluate the performance of rough-fuzzy clustering algorithm incorporating the concepts of rough sets [15].

The α index is defined in (1)
where

$$\alpha = \frac{1}{c}\sum_{i=1}^{c}\frac{wA_i}{wA_i + \tilde{w}B_i} \tag{1}$$

where

$$A_i = \sum_{x_j \in \underline{A}(\beta_i)} (\mu_{ij})^{m_1} = |\underline{A}(\beta_i)|\ ;\ \text{and}\ B_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1} \tag{2}$$

In (2) μ_{ij} represents the probabilistic memberships of object x_j in cluster β_i. The parameters w and $\overline{w}$ correspond to the relative importance of the lower and boundary regions, respectively.

The α index represents the average accuracy of c clusters. It is the average of the ratio of the number of objects in lower approximation to that in upper approximation of each cluster. In effect, it captures the average degree of completeness of knowledge about all clusters. A good clustering procedure should make all objects as similar to their centroids as possible. The α index increases with an increase in similarity within a cluster. Therefore, for a given data set and c value, the higher the similarity values within the clusters, the higher would be the α value. The value of α also increases with c. In an extreme case when the number of clusters is maximum, i.e., $c = n$, the total number of objects in the data set, the value of $\alpha = 1$. When $\overline{A}(\beta_i) = \underline{A}(\beta_i), \forall i$, that is, all the clusters $\{\beta i\}$ are exact or definable, then we have $\alpha = 1$ 3. Whereas if $\overline{A}(\beta_i) = B(\beta_i), \forall i$, the value of $\alpha = 0$. Thus, $0 \leq \alpha \leq 1$.

The ρ index represents the average roughness of c clusters and is defined in (3) by subtracting the average accuracy α from 1.

$$\rho = 1 - \alpha = 1 - \frac{1}{c}\sum_{i=1}^{c} \frac{wA_i}{wA_i + \tilde{w}B_i} \tag{3}$$

where A_i and B_i are given by Equation 2. Note that the lower the value of ρ, the better is the overall clusters approximations. Also, $0 \leq \rho \leq 1$. Basically, ρ index represents the average degree of incompleteness of knowledge about all clusters.

The α^* index is defined in (4)

$$\alpha^* = \frac{C}{D}; \quad \text{where } C = \sum_{i=1}^{c} wA_i; \quad \text{and } D = \sum_{i=1}^{c} \{wA_i + \tilde{w}B_i\} \tag{4}$$

where A_i and B_i are given by Equation 2. The α^* index represents the accuracy of approximation of all clusters. It captures the exactness of approximate clustering. A good clustering procedure should make the value of α^* as high as possible. The α^* index maximizes the exactness of approximate clustering.

The τ index is the ratio of the total number of objects in lower approximations of all clusters to the cardinality of the universe of discourse U and is given in (5)

$$\tau = \frac{R}{S}; \quad \text{where } R = \sum_{i=1}^{c} |\underline{A}(\beta_i)|; \quad \text{and } S = |U| = n. \tag{5}$$

The τ index basically represents the quality *of* approximation of a clustering algorithm.

7 Rough Fuzzy C-Medoids and Amino Acid Sequence Analysis

In most pattern recognition algorithms, symbolic representation of amino acids cannot be used directly as input since they are non-numerical variables. They, therefore, need encoding prior to input. In this regard, a bio-basis function maps a non-numerical sequence space to a numerical feature space. It uses a kernel function to transform biological sequences to feature vectors directly. Bio-bases consist of sections of biological sequences that code for a feature of interest in the study and are responsible for the transformation of biological data to a high-dimensional feature space. Transformation of input data to a high-dimensional feature space is performed based on the similarity of an input sequence to a bio-basis with reference to a biological similarity matrix. Thus, the biological content in the sequences can be maximally utilized for accurate modeling. The use of similarity matrices to map features allows the bio-basis function to analyze biological sequences without the need for encoding.

One of the important issues for the bio-basis function is how to select the *minimum set of bio-bases with maximum information*. Here, we present an application of the rough-fuzzy clustering algorithms where the c centroids mean

for c medoids, i.e., we use rough-fuzzy c-medoids (RFCMdd) algorithm [16] to select the most informative bio-bases. The objective of the RFCMdd algorithm for selection of bio-bases is to assign all amino acid subsequences to different clusters. Each of the clusters is represented by a bio-basis, which is the medoid for that cluster. The process begins by randomly choosing a desired number of subsequences as the bio-bases. The subsequences are assigned to one of the clusters based on the maximum value of the similarity between the subsequence and the bio-basis. After the assignment of all the subsequences to various clusters, the new bio-bases are modified accordingly.

The performance of RFCMdd algorithm for bio-basis selection is presented using five whole human immunodeficiency virus (HIV) protein sequences and Cai-Chou HIV data set, which can be downloaded from the National Center for Biotechnology Information [20]. The performance of different c-medoids algorithms such as hard c-medoids (HCMdd), fuzzy c-medoids (FCMdd) [11], rough c-medoids (RCMdd)[16], and rough-fuzzy c-medoids (RFCMdd) [16] is reported with respect to β index and γ index in [16]. Some of the results (shown in Fig. 10) establish the superiority of RFCMdd with lowest γ index and highest β index. Here β index signifies the average normalized homology alignment scores of input sub-sequences with respect to their corresponding medoids or bio-bases. That is, β, providing a measure of homology alignment score within a cluster, should be as high as possible. γ index, on the other hand, provides maximum normalized homology alignment score between all bio-bases, and therefore a low value is desirable. Note that homology alignment score between a pair of two sequences of amino acids measures the similarity between them in terms of probability of mutation of two amino acids, as computed from 20×20 Dayhoff mutation matrix [4].

8 Rough Ensemble Classifier for Web Services

In the previous sections we have explained the concept of knowledge encoding using rough sets, rough-fuzzy approach for modeling the concept of f-granulation of CTP, and have shown, as an example, how fuzzy granular computation provides a case generation method for decision making which is efficient in terms of classification performance, retrieval time and feature storage. Apart from granulation, the capability of rough sets (in terms of lower and upper approximations) for determining exactness in class definition for ambiguous regions is explained. Its merits for both hard and fuzzy clustering are also illustrated for brain image segmentation and determination of bio-bases from protein sequences. It is shown that incorporation of rough sets make rough-fuzzy clustering faster that fuzzy clustering.

In the present section, we describe another application of rough sets as an ensemble classifier with characteristic features. Here the problem of web service classification is considered as an example to demonstrate its efficiency through various experimental results. (The concept may be extended further into rough-fuzzy framework considering the classes and granular fuzzy either singly or together depending on the problem domain.)

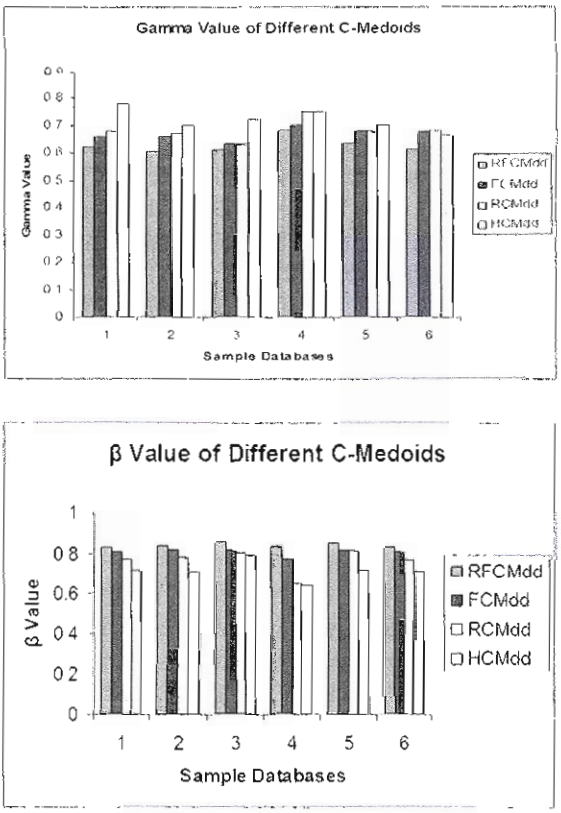

Fig. 10. β and γ values corresponding to different c-medoids (i.e., c-biobases) for different data bases

Rough Ensemble Classifier

In the problem of classification we train a learning algorithm and validate the trained algorithm. This task is performed, using some test-train split on a given labeled dataset. In the notion of rough set, let U be the given categorized dataset and $P = C_1, C_2, \ldots, C_k$ where $C_i \leq \varphi$ for $i = 1, 2, 3, \ldots, k$, $\bigcup_{i=1}^{k} C_i = U$ and $C_i \cap C_j = \varphi$ for $i \neq j$ and $i, j = 1, 2, 3, \ldots, k$, be a partition on U which provides the given k categories of U. Output of a classifier determines a new partition on U. In rough set terminology each class of the given partition P is a given concept about dataset and output of a classifiers determines new concepts about the same dataset. The given concepts can be expressed approximately by upper and lower approximations constructed by generated concepts.

The rough ensemble classifier is designed to extract decision rules from trained classifier ensembles that perform classification tasks [36]. The classification method (RSM) utilizes trained ensembles to generate a number of instances consisting of prediction of individual classifiers as conditional attribute values

and actual classes as decision attribute values. Then a decision table is constructed using all the instances with one instance in each row. Once the decision table is constructed, rough set attribute reduction is performed to determine core and minimal reducts. The classifiers corresponding to a minimal reduct are then taken to form classifier ensemble for RSM classification system. From the minimal reduct, the decision rules are computed by finding mapping between decision attribute and conditional attributes. These decision rules obtained by rough set technique are then used to perform classification tasks. The following theorems exist in this regard.

Theorem 1. *The rough set based combination of classifiers provides an optimal classifier combination technique[36].*

Theorem 2. *The performance of the rough set based ensemble classifier is at least same as that of every one of its constituent single classifiers [36].*

Some of the experimental results are shown in Fig. 11 to evaluate the performance of RSM, especially in comparison to other methods for combining classifiers, such as bagging, boosting, voting and stacking. Five learning algorithms have been used in the base-level experiments: tree-learning algorithm C4.5, the rule-learning algorithm CN2, the k-nearest neighbor (k-NN) algorithm, support vector machine (SVM), and naive bayes method. Data sets used are Reuters [6], 20NG, webKB [7], Dmoz [8], and Looksmart [9].

On each of the five text corpus, RSM as shown in Fig. 11, is found to perform better [36] than the other three classification techniques, namely Adaboost, Bagging and Stacking.

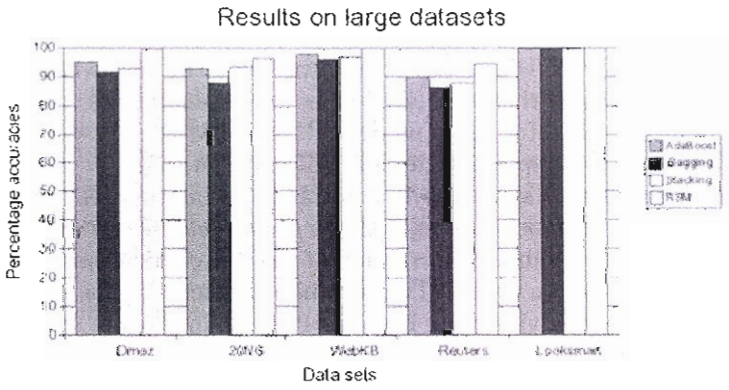

Fig. 11. Accuracy comparison of rough set based ensemble classifier with other ensemble classifiers on large datasets

Web Service Classification

The transition of the World Wide Web from a paradigm of static Web pages to one of dynamic Web services raises a new and challenging problem of locating desired web services. With the expected growth of the number of Web services available on the web, the need for mechanisms that enable the automatic categorization to organize this vast amount of data, becomes important.

A major limitation of the Web services technology is that finding and composing services require manual effort. This becomes a serious burden with the increasing number of Web services. Describing and organizing this vast amount of resources is essential for realizing the web as an effective information resource. Web Service classification has become an important tool for helping discovery and integration process to organize this vast amount of data. For instance, for categorization in the UDDI (Universal Description Discovery and Integration) registry, one needs to divide the publicly available Web Services into a number of categories for the users to limit the search scope. Moreover, Web Services classification helps the developer to build integrated Web Services. Traditionally, Web Service classification is performed manually by domain experts. However, human classification is unlikely to keep pace with the rate of growth of the number of Web Services. Hence, as the web continues to increase, the importance of automatic Web Service classification becomes necessary.

We treat the determination of a web services category as a tag-based text classification problem, where the text comes from different tags of the WSDL file and from UDDI text. Unlike standard texts, WSDL (Web Services Description Language) descriptions are highly structured. We therefore provide a Tensor space model (TSM) [37] for data representation and rough set based approach [36] for the classification of Web services. A WSDL page can be better represented as a Tensor (i.e., set of vectors corresponding to different vector spaces representing the different tags of WSDL pages). In other words, tag based TSM for web services consists of a two dimensional tensor where one dimension represents the tags of WSDL and the other represents the terms extracted from WSDL. (Unlike matrix, the number of items corresponding to tags are different.) As compared to vector space model (VSM), TSM has less complexity and captures the structural representation of WSDL better [37].

Therefore, the tensor space model captures the information from internal structure of WSDL documents along with the corresponding text content. Rough sets are used here to combine information of the individual tensor components for providing classification results. Two-step improvements on the existing classification results of web services have been shown here. In the first step we achieve better classification results over existing, by using proposed tensor space model. In the second step further improvement of the results has been obtained by using Rough set based ensemble classifier [36]. The experimental results demonstrate that splitting the feature set based on structure improves the performance of a learning classifier. By combining different classifiers it is possible to improve the performance even further.

We gathered corpuses of web services from SALCentral and webservicelist, two categorized web service indexes. The actual taxonomy exists in the web service indexes consists of more classes organized in a tree structure, but in order to simplify the task for our experiment we used only the classes from the taxonomy that were direct descendants of the root. We then discarded categories with less than ten instances, remaining categories have been used in our experiments. The discarded web services tended to be quite obscure, such as a search tool for a music teacher in an area specified by ZIP code. Details of the corpuses have been given below.

Salcentral dataset: Business-22, Communication-44, Converter-43, Country Info-62, Developers-34, Finder-44, Games-42, Mathematics-10, Money-54, News-30, Web-39.

Webservicelist dataset: Access & Security-27, Address / Locations-57, Business & Finance-97, Developer Tools-54, Content & Databases-24, Politics & Government-56, Online Validations-26, Stock Quotes-31, Search & Finders-22, Sales Automation-20, Retail Services-30.

Two-step improvement on the existing classification results of web services has been shown. In the first step we achieve better classification results by using the proposed tensor space model. In the second step further improvement of the results has been obtained by using the Rough set based classifier. In Table 1 percentage accuracies of classifications have been compared on two different representation models.

WSDL documents corresponding to above datasets have been represented in vector space model and tag based tensor space model, respectively [37]. Three well known classifiers, namely, naive bayes (NB), namely, support vector machine (SVM) and decision tree (C4.5) have been considered to provide classification results in two different models [37]. Note that classification results in tensor space model have been computed on individual tensor components and combining them with majority voting. The results show that classification on tensor space model provides better percentage accuracy than vector space model for both the datasets and for all classifiers considered.

In Fig. 12, bar chart of percentage accuracies of combining classifiers have been given. Here all the classifiers have been tested on meta data generated by the base level classifiers from each tensor component. Naive bayes (NB), support vector machine (SVM), decision tree (DT), majority vote and rough set

Table 1. VSM vs. TSM

Data	Model	NB	SVM	C4.5
salcentral	VSM	47.06	55.50	49.62
salcentral	TSM	53.45	58.78	56.31
webservicelist	VSM	62.21	65.57	64.19
webservicelist	TSM	69.48	73.63	70.90

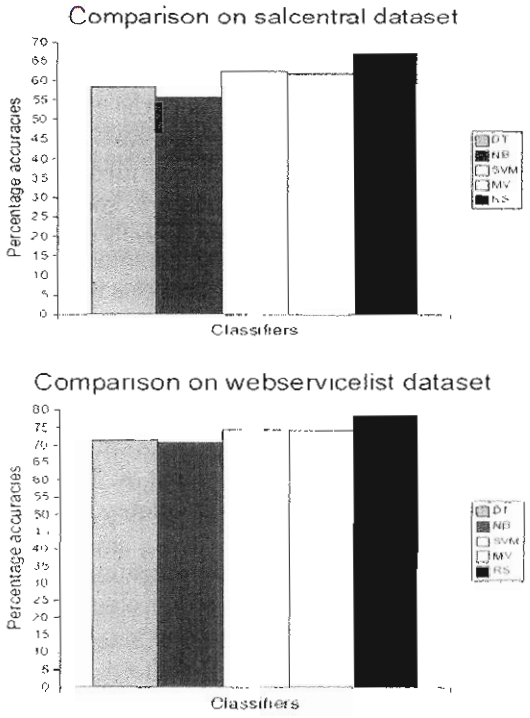

Fig. 12. Comparison of percentage accuracies of classifiers on salcentral and webservicelist datasets

(RS) are applied to combine the output of base level classifiers corresponding to individual tensor components. Results show that, rough set provides the better classification results than other methods on both datasets considered.

9 Conclusions

The concepts of knowledge encoding using rough sets, judicious integration of the merits of rough and fuzzy sets, rough-fuzzy granulation, and their relevance to computational theory of perception (CTP) are explained. Significance of granular computing and the formulation of rough ensemble classifier are illustrated. Ways of modeling f-granulation property of CTP are discussed. Three examples of judicious integration, viz., rough-fuzzy case generation, rough-fuzzy c-means and rough-fuzzy c-medoids are explained along with their merits and some quantitative indices. These rough-fuzzy methodologies can be viewed under the generalized rough set theoretic framework with four cases, namely, (i) crisp set crisp granules, (ii) crisp set fuzzy granules, (iii) fuzzy set crisp granules, and (iv) fuzzy set fuzzy granules. For example, the ensemble classifier concerns with case (i) whereas case-based reasoning and clustering methods belong to cases (ii) and (iii) respectively.

Significance of rough-fuzzy clustering in protein sequence analysis for determining bio-bases and segmentation of brain MR images is explained. Effectiveness of rough ensemble classifier, which provides an optimum combination, is demonstrated in web service classification. Concept of fuzzy granulation through rough-fuzzy computing, and performing operations on fuzzy granules provide both information compression and gain in computation time; thereby making it suitable for data mining applications. Other application of fuzzy information measures in fuzzy approximation space for feature selection problem is recently reported in [17]. The concept of rough ensemble classifier can be extended into rough-fuzzy framework with appropriate criterion depending on the application domain.

Rough-fuzzy granular computing (GrC), coupled with computational theory of perception (CTP), have great promise for human like decision making and efficient mining of large, heterogeneous data, and providing solution of various real-life ambiguous recognition problems. Apart from the problem of defining granules with their sizes appropriately, future challenges in GrC and CTP, include formulating efficient methodologies based on fuzzy *granular computing* and granular*fuzzy computing* for making the aforesaid tasks of decision making more natural and efficient. While the former concerns with computation using fuzzy granules, the latter deals with fuzzy computing using granules. Another important application of fuzzy GrC would be granular information retrieval from heterogeneous media like WWW.

Acknowledgement

The author acknowledges J.C. Bose Fellowship of the Govt. of India, as well as his co-investigators Dr. C.A. Murthy, Dr. P. Mitra, Dr. P. Maji, Mr. S. Saha, and Mr. D. Sen.

References

1. Baldi, P., Brunak, S.: Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge (1998)
2. Banerjee, M., Mitra, S., Pal, S.K.: Rough Fuzzy MLP: Knowledge Encoding and Classification. IEEE Trans. Neural Networks 9, 1203–1216 (1998)
3. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithm. Plenum, New York (1981)
4. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: A Model of Evolutionary Change in Proteins. Matrices for Detecting Distant Relationships, Atlas of Protein Sequence and Structure 5, 345–358 (1978)
5. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2005)
6. http://www.daviddlewis.com/resources/testcollections/reuters21578/
7. http://www.cs.cmu.edu/~WebKB/
8. http://www.dmoz.org/
9. http://www.looksmart.com

10. Kolodner, J.L.: Case-Based Reasoning. Morgan Kaufmann, San Mateo (1993)
11. Krishnapuram, R., Joshi, A., Nasraoui, O., Yi, L.: Low complexity fuzzy relational clustering algorithms for web mining. IEEE Trans. on Fuzzy System 9, 595-607 (2001)
12. Kuipers, B.J.: Qualitative Reasoning. MIT Press, Cambridge (1984)
13. Li, Y., Shiu, S.C.K., Pal, S.K.: Combining feature reduction and case selection in building CBR classifiers. IEEE Trans. on Knowledge and Data Engineering 18, 415-429 (2006)
14. Lingras, P., West, C.: Interval set clustering of web users with rough K-means. Journal of Intelligent Information Systems 23, 5-16 (2004)
15. Maji, P., Pal, S.K.: Rough set based generalized fuzzy C-means algorithm and quantitative indices. IEEE Trans. on System, Man and Cybernetics, Part B, 37, 1529–1540 (2007)
16. Maji, P., Pal, S.K.: Rough-fuzzy C-medoids algorithm and selection of bio-basis for amino acid sequence analysis. IEEE Trans. Knowledge and Data Engineering 19, 859-872 (2007)
17. Maji, P., Pal, S.K.: Feature Selection Using f-Information Measures in Fuzzy Approximation Spaces. IEEE Trans. Knowledge and Data Engineering (to appear)
18. Mitra, S., Banka, H., Pedrycz, W.: Rough-fuzzy collaborative clustering. IEEE Trans. on Systems, Man, and Cybernetics - Part B: Cybernetics 36, 795-805 (2006)
19. Mitra, S., De, R.K., Pal, S.K.: Knowledge Based Fuzzy MLP for Classification and Rule Generation. IEEE Trans. Neural Networks 8, 1338-1350 (1997)
20. National Center for Biotechnology Information, `http://www.ncbi.nlm.nih.gov`
21. Pal, S.K., Talwar, V., Mitra, P.: Web mining in soft computing framework: Relevance, state of the art and future directions. IEEE Trans. Neural Networks 13, 1163-1177 (2002)
22. Pal, S.K., Bandyopadhyay, S., Ray, S.S.: Evolutionary Computation in Bioinformatics: A Review. IEEE Transactions on Systems, Man, and Cybernetics, Part-C 36, 601-615 (2006)
23. Pal, S.K., Skowron, A. (eds.): Rough-Fuzzy Hybridization: A New Trend in Decision Making. Springer, Singapore (1999)
24. Pal, S.K., Polkowski, L., Skowron, A. (eds.): Rough-neuro Computing: A Way to Computing with Words. Springer, Berlin (2003)
25. Pal, S.K., Skowron, A. (eds.): Special issue on Rough Sets, Pattern Recognition and Data Mining. Pattern Recognition Letters 24 (2003)
26. Pal, S.K., Dillon, T.S., Yeung, D.S. (eds.): Soft Computing in Case Based Reasoning. Springer, London (2001)
27. Pal, S.K., Shiu, S.C.K.: Foundations of Soft Case Based Reasoning. John Wiley, NY (2003)
28. Pal, S.K., Mitra, P.: Case generation using rough sets with fuzzy discretization IEEE Trans. Knowledge and Data Engineering 16, 292-300 (2004)
29. Pal, S.K., Mitra, P.: Pattern Recognition Algorithms for Data Mining. Chapman & Hall CRC Press, Boca Raton (2004)
30. Pal, S.K., Ghosh, A., Sankar, B.U.: Segmentation of remotely sensed images with fuzzy thresholding, and quantitative evaluation. International Journal of Remote Sensing 2, 2269-2300 (2000)
31. Pal, S.K., Polkowski, L., Skowron, A. (eds.): Rough-Neural Computing Techniques for Computing with Words. Springer, Berlin (2004)
32. Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.): PReMI 2005. LNCS, vol. 3776. Springer, Heidelberg (2005)

33. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic, Dordrecht (1991)
34. Pedrycz, W., Skowron, A., Kreinovich, V. (eds.): Handbook of Granular Computing. John Wiley, N.Y. (2008)
35. Ray, S.S., Bandyopadhyay, S., Mitra, P., Pal, S.K.: Bioinformatics in Neurocomputing Framework. IEE Proc. Circuits Devices & Systems 152, 556–564 (2005)
36. Saha, S., Murthy, C.A., Pal, S.K.: Rough set Based Ensemble Classifier for Web Page Classification. Fundamentae Informetica 76, 171–187 (2007)
37. Saha, S., Murthy, C.A., Pal, S.K.: Classification of Web Services using Tensor Space Model and Rough Ensemble Classifier. In: Proc. 17th International Symposium on Methodologies for Intelligent Systems, Toronto, Canada, pp. 508–513 (2008)
38. Sen, D., Pal, S.K.: Histogram Thresholding using Fuzzy and Rough Measures of Association Error. IEEE Trans. Image Processing 18, 879–888 (2009)
39. Sen, D., Pal, S.K.: Generalized Rough Sets, Entropy and Image Ambiguity Measures. IEEE Trans. Syst, Man and Cyberns. Part B 39, 117–128 (2009)
40. Sun, R.: Integrating Rules and Connectionism for Robust Commonsense Reasoning. Wiley, N.Y. (1994)
41. Swiniarski, R.W., Skowron, A.: Rough Set Methods in Feature Selection and Recognition. Pattern Recognition Letters 24, 833–849 (2003)
42. Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.): RSKT 2008. LNCS (LNAI), vol. 5009. Springer, Heidelberg (2008)
43. Yao, Y.Y.: Granular Computing: Basic Issues and Possible Solutions. In: Proceedings of the 5th Joint Conference on Information Sciences, vol. I, pp. 186–189 (2000)
44. Zadeh, L.A.: A new direction in AI: Toward a computational theory of perceptions. AI Magazine 22, 73–84 (2001)
45. Zadeh, L.A., Pal, S.K., Mitra, S.: Foreword. In: Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing. Wiley, New York (1999)
46. Zadeh, L.A.: Fuzzy Logic, Neural Networks, and Soft Computing. ACM 37, 77–84 (1994)
47. Zadeh, L.A.: Towards a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic. Fuzzy Sets Systems 19, 111–127 (1997)