Granular Data Mining:
Concepts, Models and Applications

Sankar K. Pal
Indian Statistical Institute
Calcutta
http://www.isical.ac.in/~sankar
Contents

- Pattern Recognition & DM
- Granular Computing
  - Granulation & Granules
  - Fuzzy sets and uncertainty analysis
  - Rough sets and information granules
  - Example: Case mining
- Generalized Rough Sets and Entropy
- Video Processing: Object extraction and tracking
- Bioinformatics: miRNA selection
- Social Networks: Community detection
- Challenging issues
- Relevance to Big Data
PR Tasks & Challenges

- **Classification:** Sampled data are given about the pattern space. And the Challenge is to estimate the unknown regions of the pattern space based on the sampled data (incomplete information) → **Abstraction + Generalization**

- **Clustering:** Entire data is given. And the Challenge is to partition it into meaningful regions. Number of regions may be known or unknown
Image Classification \(\rightarrow\) Pixel Classification \(\rightarrow\) Supervised

Image Segmentation \(\rightarrow\) Pixel clustering \(\rightarrow\) Unsupervised
Pattern Recognition and Machine Learning principles applied to a very large (both in size and dimension) heterogeneous database

≡ Data Mining

Data Mining + Knowledge Interpretation

≡ Knowledge Discovery

Process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data
Data Science

- Concerns with the acquisition, storage, retrieval, processing, mining and finally the conversion of data into knowledge.

- Employs techniques and theories drawn from many fields within the broad areas of e.g., mathematics, statistics, information science, and computer science.

- Although, it is not necessarily restricted to Big data, methods that are scalable to big data are of much interest to data science.
Granular Computing?

- Granulation
- Granules
- Computing with granules + Uncertainty handling
- Relevance of Rough sets and fuzzy sets
Granulation: Natural clustering

- Replacing a fine-grained universe by a coarse-grained one, more in line with human perception
- A process like - self-organization, self-production, morphogenesis, Darwinian evolution – that are abstracted from natural phenomena
- Clusters/segments so formed by granulation (natural clustering) are called Granules
- Process of formation and representation of Granules
Concept of Granules

- Clump of indiscernible entities (drawn together by similarity, proximity or functionality) w.r.t. given attributes

- Examples: Granules in
  - **Age**: very young, young, not so old, ...
  - **Direction**: slightly left, sharp right, ...
  - **School**: each class/section
  - **Image**: regions of similar colors, gray values
    e.g., max diff of 6 gray levels (Weber’s law)

- Granules may be Crisp or Fuzzy (overlapping)
Depending on size and shape, granules have -

- different levels of granularity,
- characterize a specific aspect of the problem
- represent the model differently and
- regulate the decision-making tasks accordingly

Granules play a significant role in GrC
Granular Computing (GrC): An information processing paradigm - that works with the process of *information granulation*/*abstraction*, and - where computation is performed using *information granules* and not the data points (objects)

- Information compression
- Computational gain

- Suitable for Mining Large Data
Concept of granulation is inherent in theories of Fuzzy Sets & Rough Sets

- Membership function, building block in FS, practically granulates the features; producing fuzzy granulation of feature space

- Rough set theory concerns with a granulated domain (crisp set defined over a crisp granulated domain)

FS and RS most successful technologies for GrC
Fuzzy Sets
Classical Set $\mu \in \{0,1\}$ Hard

Fuzzy Set $\mu \in [0,1]$ Soft

Examples: tall man, long street, large number, sharp corner, very young, etc.

- FS are nothing but $\mu$-functions
- $\mu$-functions are context dependent

$A = \{ (\mu_A(x), x) : \text{for all } x \in X \}$

$\mu_A(x)$: Membership function: degree of belonging of $x$ to $A$ or degree of possessing some imprecise property represented by $A$
Fuzzy Sets are nothing but Membership Functions
Membership Function: Context Dependent
Characteristics of FS

- FS is a Generalization of classical set theory
  - Greater flexibility in capturing faithfully various aspects of incompleteness or imperfection in a situation

- Flexibility is associated with the Concept of $\mu$
  - As $\mu$ ↑ Amount of Stretching the Concept ↓
  - FS are Elastic, Hard sets are inelastic

- $p(x)$: Concerns with the no. of occurrences of $x$
- $\mu(x)$: Concerns with the compatibility (similarity) of $x$ with an imprecise concept
Concept of Flexibility & Uncertainty Analysis (overlapping data/ concept/ regions)
Set (crisp) defined over Granulated (crisp) domain
\[ [x]_B = \text{set of all points belonging to the same granule as of the point } x \text{ in feature space } \Omega_B. \]

\[ [x]_B \text{ is the set of all points which are indiscernible with point } x \text{ in terms of feature subset } B. \]
Approximations of the set \( X \subseteq U \) w.r.t feature subset B

**B-lower:** \( B\bar{X} = \{x \in U : [x]_B \subseteq X\} \) \hspace{1em} Granules definitely belonging to \( X \)

**B-upper:** \( \bar{B}X = \{x \in U : [x]_B \cap X \neq \emptyset\} \) \hspace{1em} Granules definitely and possibly belonging to \( X \)

If \( B\bar{X} = \bar{B}X \), \( X \) is \( B \)-exact or \( B \)-definable

Otherwise it is \textit{Roughly definable}

Rough Sets are Crisp Sets, but with rough description
Rough Sets

Uncertainty Handling
(Using lower & upper approximations)

Granular Computing
(Using information granules)

Two Important Characteristics
What is Lower & Upper Approximations of Clusters and Uncertainty Modelling?
Cluster definition using rough lower & upper approx

- Sets and Granules can either or both be fuzzy (in real life)
- Lower and upper approximate regions could be fuzzy \((m.func)\)
- Generalized Rough Sets – Stronger model of uncertainty handling
  (uncertainty due to overlapping regions + granularity in domain)

\[
\text{Roughness in } \beta : 1 - \frac{|\text{lower}|}{|\text{upper}|}
\]
Before I describe the Generalized Rough Sets and example applications of Granular Mining, let me explain the

- Concept of granulation & granules
- f-information granules & case mining
- Applications of (rough) information granules
  - Relevance of Rough-Fuzzy computing in SC paradigm
Concept of -

$f$- Information Granules using Rough Rules
- Elongated cluster
- Case generation of reduced dim

• Rule provides crude description of the class using granule

\[ M_1 \land M_2 \]
Rule characterizing the granule can be viewed as the Case or Prototype representing the class/concept/region.

Elongated objects need multiple rules/granules.

Unsupervised: No. of granules is determined automatically.

Cases (prototypes) are granules, not sample points case generation, NOT selection.
All the features may not appear in rules

- Dimensionality reduction

Depending on topology, granules of different classes may have different dimensions

- Variable dimension reduction

- Less storage requirement

- Fast retrieval

Suitable for mining data with large dimension and size
Example: IRIS data case generation

Three flowers: Setosa, Versicolor and Virginica
No of samples: 50 from each class
Features: sepal length, sepal width, petal length, petal width
Iris Followers: Setosa, Versicolor and Virginica

(a) Sepal L - Sep W
(b) Sepal L – Petal L
(c) Sepal L – Petal W
Iris Followers: Setosa, Versicolor & Virginica

(a) Petal L - Sepal W
(b) Petal W - Sepal W
(c) Petal W - Petal L
Iris Flowers: 4 features, 3 classes, 150 samples

Number of cases = 3 (for all methods)
Applications of Rough Information Granules

- Case Based Reasoning (evident is sparse)
- Prototype generation and class representation
- Clustering & Image segmentation (k selected autom)
- Case representation and indexing
- Knowledge encoding
- Dimensionality reduction
- Data compression and storing
- Granular information retrieval
Granular Computing (GrC): Computation is performed using *information granules* and not the data points (objects)

- Information compression
- Computational gain

➢ Suitable for Mining Large Data

- Rough set theory enriched GrC research
Concept of -

Generalized Rough Sets
Uncertainty modeling)

Incorporate fuzziness in set & granules of rough sets

Fuzzy lower and upper approximate regions
(characterized by fuzzy membership functions.)
Generalized Rough Sets

In practice, the Set and Granules, either or both, could be Fuzzy.

- Generalized Rough Set
- Stronger Paradigm for Uncertainty Handling
Incorporate Fuzziness in Set & Granules

- $X$ is a **crisp** set & Granules have **crisp** boundaries - rough set of $X$ (**crisp** low-upp memb func)
- $X$ is a **fuzzy** set & Granules have **crisp** boundaries - rough-fuzzy set of $X$ (**fuzzy** low-upp memb func)
- $X$ is a **crisp** set & Granules have **fuzzy** boundaries — fuzzy-rough set of $X$ (**fuzzy** low-upp memb func)
- $X$ is a **fuzzy** set & Granules have **fuzzy** boundaries — fuzzy rough-fuzzy set of $X$ (**fuzzy** low-upp memb func)

Generalized Rough Sets
Generalized Rough Sets

When $R$ is an equivalence relation

$X$ is a **crisp** set & Granules have **crisp** boundaries

$$RX = \{ u \mid u \in U : [u]_R \subseteq X \}$$

$$\overline{RX} = \{ u \mid u \in U : [u]_R \cap X \neq \emptyset \}$$

$[u]_R$ represents the granule that contains $u$.

The pair $< RX, \overline{RX} >$ is referred to as the rough set of $X$. 
R is an equivalence relation

\[ X \] is a fuzzy set & Granules have fuzzy boundaries

The pair \( <RX, \bar{RX}> \) is referred to as the fuzzy rough-fuzzy set of \( X \).
**Roughness Measure**

\[ \rho_R(X) = 1 - \frac{|RX|}{|RX|} \]

- a measure of inexactness of \( X \)
- \( RX \) and \( \overline{RX} \) are the lower and upper approxs. of \( X \)
Plots of entropy for different values of base $\beta$ and gain functions

$e = 2.718$

\[ A \rightarrow \rho_R(X) \]

\[ B \rightarrow \rho_R(X^C) \]
Several Applications in Data Analysis

Example: Image Analysis

- R-F entropy takes care of fuzzy boundaries of regions + rough resemblance between nearby gray levels + rough resemblance between nearby pixels (i.e., fuzziness + granulation)

Nearby gray levels have limited discernibility

Granule Example: A region containing gray values separated by 6 gray levels. (Weber’s law)
Example Applications of Granular Mining

- Video tracking
- Gene and miRNA selection (Bioinformatics)
- Community detection (Social networks)

Role of -
- Granules
- Lower approximation
- \( f \)-information measure
Example:

Generalized Rough-fuzzy Image Entropy and Tracking

- Granules determine *Roughly Resemble* gray levels & pixels
- Principle: *Minimizing Entropy* (→ maximizing O/B separation)
  - Effect of granules in object extraction
    - Fuzzy image entropy vs. Rough-fuzzy image entropy
    - Granules of equal size vs. unequal size vs. arbitrary shape/size
1-d and 2-d Granules for computing GA and SA

Example:
Segmentation Results

- **Effect of granules**
  \( (\omega = 6) \)

- **Baboon image**
- **Proposed r-f entropy**
- **Fuzzy entropy**

- **Brain MR image**
- **Proposed r-f entropy**
- **Fuzzy entropy**

- **Remote sensing image**
- **Proposed r-f entropy**
- **Fuzzy entropy**
So far we considered granules of Equal size.

Next, consider granules of Unequal size:

- *Quad-tree decomposition* (spatial)
- Spatial image segmentation
Example: Quad-tree decomposition and granule formation
Example Comparison

Original

Otsu’s thresholding

RE with 4x4 granule

RE with 6x6 granule

Rough-fuzzy with crisp set and 6x6 granule

Proposed methodology
Video Tracking

- Spatial segmentation on each frame
- Temporal segmentation based on 3 previous frames

- Merits of unequal size over fixed size granules
  (Tracker signifies segmented output)
Other RE (6x6) - SP

Proposed

Otsu

RE (6x6) - PUM

RE (4x4) - PUM
Variation of $\beta$-Index over sequence ‘a’

- Homogeneous granules of unequal size reduce the formation of spurious segments → Reduce abrupt change of index-value over frames.
Granules of Arbitrary Shape (natural)

Use Neighborhood Granules in RGB-D:
3-d spatio-temporal, 2-d spatio-color, 1-d color

- Granules obtained from *Lower Approx. Object Model*
- Use granules, instead of pixels, for O/B partition
- *Granular level* rule based decision
- Automatic updation of rule base with flow graph

Deals with ambiguous tracking situation (unsupervised)
(overlapping, *newly appeared object*, merged with similar color)
3-D temporal information in Video ($P = 4$)

- Value at point $(m, n, p)$ of the changed information matrix in D-space:
  
  $(3^{rd}$ dim represents change in information)

  $P = 1, 2, ..., P-1$

- Diff is computed w.r.t. depth value only

\[
\tau_p(m, n) = |f_t(m, n) - f_{t-p}(m, n)|
\]
What is Lower Approximation Object Model?
Extraction of Temporal Information (in D-space, $P = 4$)

Current Frame $f_t$

$|f_t - f_{t-1}| = |f_t - f_{t-3}|$: Lower & Upper approx. of object
Granules extracted in lower approx object model in $\tau_{med}$ frame:

- 2-d spatio color
- 1-d color
- 3-d spatio temp

$$\tau_p = |f_i - f_{p-1}|$$

Representative location of a 2D spatio-color granule

$\tau_{med} \quad \tau_p \quad \ldots \ldots \ldots \ldots \ldots \tau_1$
Example:
Spatio-color Granules
(in RGB-space)
Methodology

- Initial P frames
- Unsupervised Initial Labelling
- $f_t \cdot f_{t-\Delta}$
- Spatio-temporal Granulation
- 3-D color granule (RGB) depth granule (D) and temporal granule
- Compute 1-d color granule (RGB) depth granule (D) and temporal granule
- Rule-Base
- Segment Foreground
- Compute the Coverage of Rule-base
- Obtained Object Model
- Update Rule-Base Using Granular Flow Graph
- Form Granular Flow Graph
- Is it Satisfactory?

Yes
- Track the Segment

No
- Compute the Coverage of Rule-base
Overlapping objects  (with Granulation + FL Adaptation)
Without Flow Graph Based Adaptation

Newly appeared object: Fails
With Flow Graph Based Adaptation

Newly appeared object: Succeeds
Multiple moving objects (with Granulation + FL Adaptation)

Frames per sec = 15, P = 6
Another Application on –

Relevance of Lower Approximation
Example:

miRNA Ranking in Cancer Detection

- Small sample, large dimension
- Set is crisp (C or N) & Granules are fuzzy: *Fuzzy-rough entropy*
- *Fuzzy Lower Approx.* of N & C classes: used to find prob. (relative frequency) of definite and doubtful regions for entropy computation

- Entropy minimization implies higher *Relevance* of a miRNA
- Top 1% miRNAs provides significant improvement over entire set in terms of F-score
- Superiority over related methods
Crisp Classes & Fuzzy Granules of Patients

Cancer Class (C)

Normal Class (N)

Normal patient

Normal granule

Cancer patient

Cancer granule
Granulation w.r.t. miRNA-M1 (1-d granules)

\[ \mu_{lowC}(p) = \min(\mu_N(p), 1), \quad p \in C \]
\[ = \min(\mu_N(p), 0), \quad p \in N \]

\[ \mu_{lowN}(p) = \min(\mu_N(p), 1), \quad p \in N \]
\[ = \min(\mu_N(p), 0), \quad p \in C \]

Rel freq Lower approx of C
\[ \lambda_{lowC} = \frac{1}{|C|} \sum_p \mu_{lowC}(p) \]

Rel freq of Lower approx of N
\[ \lambda_{lowN} = \frac{1}{|N|} \sum_p \mu_{lowN}(p) \]

\[ \mu_N(p) = \text{Membership of patient } p \text{ to Normal Class (N)} \]
\[ \mu_C(p) = \text{Membership of patient } p \text{ to Cancer Class (C)} \]
\[ \mu_N(p) + \mu_C(p) = 1 \]
\[ \lambda_1 = \text{Mean cardinality (mean relative freq.) of lower approx. regions of N & C} \]
\[ = \frac{\lambda_{\text{lowN}} + \lambda_{\text{lowC}}}{2} \]

\[ \lambda_2 = \text{Cardinality (relative freq.) of overlap region between N & C} \]
\[ = \frac{(1 - \lambda_{\text{lowN}}) + (1 - \lambda_{\text{lowC}})}{2} \]

\[ \lambda_1 + \lambda_2 = 1 \]
Entropy

- Fuzzy-Rough Entropy w.r.t. miRNA ($M_1$)

$$\text{FREM}_1 (N, C) = - \sum_{i=1}^{2} \lambda_i \log_2 \lambda_i , \quad \lambda_1 + \lambda_2 = 1$$

- $\lambda_1$: Mean cardinality of lower approx. regions of N & C
- $\lambda_2$: Cardinality of overlap region between N & C
Relevance & Redundancy

- **FREM}_i (N, C)↓**: Separability between N and C classes w.r.t. miRNA M}_i↑
  - Relevance of M}_i for cancer classification

- **Redundancy of M}_i w.r.t another miRNA M}_j**
  - Separability of M}_i w.r.t. M}_j, i ≠ j
  - Compute FREM}_i (M}_i, M}_j)

(Ex: to compute FREM}_i (M}_1, M}_2), consider M}_1 and M}_2 as two separate classes instead of N and C, and the # samples in these classes would be 7 & 7 instead of 4 & 3 respectively.)
Results: Relevance (All 1% selected)

- Classification using SVM

<table>
<thead>
<tr>
<th>miRNAs/Samples</th>
<th>Breast</th>
<th>Colorectal</th>
<th>Lung</th>
<th>Melanoma</th>
<th>Pancreas</th>
<th>Nasopharyngeal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>309/98</td>
<td>352/66</td>
<td>866/36</td>
<td>864/57</td>
<td>847/158</td>
<td>887/50</td>
</tr>
</tbody>
</table>
Example:

Fuzzy-rough Community (virtual group) in Social Networks

- High volume, dynamic, complex
- Granules model $f$-Relations/Interactions of actors
  - Output: Fuzzy-rough communities
  - Fuzzy mutual information: Measures the goodness of community structure
  - Suitable for overlapping community structure
Why Community Detection?

- In Society, one can find groups that naturally form, e.g., families, co-workers’ circles, friendship circles, villages and towns.
- Similarly, in an Online social network, we can find virtual groups, which live on the web.
- Detecting these groups (communities) has practical significance:
  - in WWW this helps to Optimize the Internet infrastructure
  - in a purchase network this can Boost the sell by Recommending appropriate products
  - in computer network it helps to Optimize the Routing table creation
  - in citation network it finds Researchers of similar fields
- Detecting these communities also helps in identifying Special actors:
  - central nodes of the clusters, or nodes in the boundary region (who act as a bridge between communities), are the special actors who play different important roles within the society.
Why Fuzzy Granules?

- A social network is viewed as a collection of –
  - Relations between social actors (nodes)
  - Interactions between social actors

- Actors are often indistinguishable in some problem solving → concept of *Granules*

- Relations/ Interactions between nodes & Clusters of nodes do not often lend themselves to precise definition → *Fuzzy boundaries*
Fuzzy Granules: Example

- Construct a granule around a node (or actor)
  - This actor is referred as the Center (Representative) of the granule
  - Radius of the granule = 1 hop distance, say

- Granule around node 1
- Granule around node 34
- Granule around node 3
Community Detection in FGSN: Principle

- **Identify the dense granules** (whose granular degree exceeds a threshold) i.e., identify $\theta$-Cores
- **Merge them if they are nearby** (search for $\theta$-Cores belonging to same community i.e., find Community reachable $\theta$-Cores)
- **Form a meaningful fuzzy community by discarding the weakly coupled granules** (whose Granular embeddedness is less than a threshold $1/\tau$)
- **Fuzzy-rough representation**

$\theta$: Density co-efficient  
$\tau$: Coupling co-efficient
Granule $A_p$ is a *directly community reachable* $\theta$-core to $A_q$, if $A_p$ and $A_q$ are both $\theta$-cores, and $A_p$ is in the neighborhood of $A_q$ i.e., (center node $p$ of $A_p$ is a support of $A_q$)

$A_p \in \Gamma(A_q)$ and $\mathcal{D}(A_q), \mathcal{D}(A_p) \geq \theta$

Representative (center) node $p$ of $A_p$ is a support of $A_q$, and therefore $A_p$ and $A_q$ are directly community reachable

→ Ensures significant amount of overlapping
Indirectly Community Reachable $\theta$-Cores

- Granule $A_p$ is an *indirectly community reachable* $\theta$-core to granule $A_q$ if –

  there exists a chain of granule centers $p_1, p_2, \ldots, p_n$ where $p_1 = p$ and $p_n = q$

  s. t. $A_{p_i}$ is directly community reachable $\theta$-core to $A_{p_{i+1}}$

For Undirected Network, $A_p$ is indirectly community reachable to $A_q$ implies $A_q$ is indirectly community reachable to $A_p$.

For Directed Network, either $A_p$ is indirectly community reachable to $A_q$ or $A_q$ is indirectly community reachable to $A_p$. 

---

- $\theta$-core
Example: Necessity to check embeddedness

$A_P$ and $A_Q$ are community reachable but they are loosely coupled (i.e., center nodes P and Q have no common neighbour)
Nodes to Community Structures
(Data to Knowledge)

- Nodes $\Rightarrow$ Granules $\Rightarrow$ $\theta$-cores $\Rightarrow$ Groups of community reachable $\theta$-cores $\Rightarrow$ Community $\Rightarrow$ Fuzzy-rough community structures
Fuzzy-Rough Community: illustration

$P_1, P_2, Q_1, Q_2$: Community reachable regions consisting of multiple $\theta$-cores

- $P_1$ and $P_2$ are merged together as they satisfy the criterion of coupling coefficient, and similarly for $Q_1$ and $Q_2$. No. of nods in the overlapping regions are 8 and 7 respectively reflecting higher coupling. This is NOT true for merging $P_1$ and $Q_1$, or $P_2$ and $Q_2$ where the nos. are 3 and 2 only reflecting lower coupling. As a result, two communities $C_1$ and $C_2$ are detected.

- Every node has a normalized membership value ($\mu$) w.r.t. the granules centered at other nodes.
- While computing the membership value ($\delta$) of a blue node for community $C_2$, say, we add up its $\mu$-values w.r.t. the granules centered ONLY at the yellow nodes. (i.e., out of 56 $\mu$-values, only 23 values are added)

$\delta_{C_1} = 1$, $\delta_{C_2} = 0$

$\delta_{C_1} = 0$, $\delta_{C_2} = 1$

0 $< \delta_{C_1}, \delta_{C_2} < 1$

They may not be equal

$C_1, C_2$: Detected communities
LFR Benchmark Data (undirected friendship network)

Network Size (nodes): 1001; Min Community Size: 150; Max Community Size: 250

\( \eta \): average fraction of edges shared with other community w.r.t. those within communities;
\( O_n \): fraction of nodes belonging to overlapped regions

- As expected, NFMI decreases in all the cases when \( \eta \) increases
- for \( \eta > 0.3 \) FRC-FGSN shows prominent improvement over other methods
- FRC-FGSN produces superior performance for \( O_n \) ranging from 0.2 to 0.4 and second best for \( O_n < 0.2 \)
- FRC-FGSN has promise for detecting overlapping communities

\( \theta = 2.0, \ 1/\tau = 0.03 \) (# communities generated 7 or 8)
Summary

- Different Machine learning tools
- Data: Videos, miRNAs and social networks

Where are these leading to?
Relevance to CTP

- *Fuzzy (F)-Granularity* characteristics of Computational Theory of Perceptions (CTP) can be modeled using *Fuzzy-Rough* computing concept

Promising Future Research Problem
Granulation is a process like self-reproduction, self-organization, functioning of brain, Darwinian evolution, group behavior, cell membranes and morphogenesis - that are abstracted from natural phenomena.

- $f$-Granulation is inherent in human thinking & reasoning process, and plays an essential role in human cognition.
Relevance to BIG Data handling

- Uncertainty handling and Granular mining points of view
Thank You!!
Giant Panda from Chendu: *Life is so..o good with bamboo shoots*