Int J Pharm Bio Sci 2015 Oct; 6(4): (B) 642 - 646

Review Article **Medical Informatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

SOFT COMPUTING TECHNIQUES FOR CATEGORICAL DATA ANALYSIS IN BIO-INFORMATICS

K.SHARMILA BANU AND B.K.TRIPATHY

SCSE, VIT University

ABSTRACT

Computer based data analytical algorithms have found immense use in all fields. Data Science is emerging as one of the prominent disciplines in Intelligent Bio-Sciences. This field requires collaborative efforts from Biological Scientist, Doctors, Epidemiologists, Computer and Data Scientists, policy makers and administrators. Huge amounts of biological, medical and epidemiological data generated are being studied for knowledge discovery and pattern mining. Knowledge gained from these data are significant as they touch human lives. Hence it is important that they are carefully stored, retrieved, analysed and mined. Categorical (non-numerical data) and high dimensional data are becoming very common in a lot of real-time Bio-Medical applications. These data are packed with information which helps users understand features better as they involve natural language in most cases. But, it is difficult to map the categorical data to scale and analyze where as it is easy in the case of continuous or numerical data. This paper discusses the features of categorical and high dimensional data as well as variants of Fuzzy and Rough set based clustering algorithms like FCM and MMR, MMeR, SDR, SSDR.

KEYWORDS: **Categorical Data, Clustering, Bio-Informatics, FCM, MMR, MMeR, SDR**

K.SHARMILA BANU
SCSE, VIT University

I. INTRODUCTION

Modern information systems for Bio-Informatics are witnessing a lot of research advancements. The nature of data involved in these systems is continuous, categorical, spatial, and evolving. Clustering such feature rich data requires robust algorithms. Crisp clustering algorithms are more suitable for continuous or numerical data. But they cannot handle multi-valued attributes and categorical data.

Table 1
An Excerpt of Data Used in [8]

Region	Species	Rainfall
Kolhapur	malbarica	1000-1100
Radhanagari	Parrotamyia	2500
Gargoti	chalarni	1400-1500
Shirol	siroi	750-800

Information systems can be constructed after data collection and preprocessing. Table 1 illustrates an excerpt of data studied in[8] Information system helps to understand the properties of biological and demographic entities. Then the data is analysed in terms of – looking for similar entities with same properties (clustering), looking for outliers (entities that deviate from expected behaviour) and so forth. Such Clustering of data can help scientists understand the properties of data and the resultant conditions. Imposing crisp boundaries on clusters may not give appropriate result and analysis. It is also observed that most of the real-time applications involve data that cannot be placed under specific boundary defined clusters[1].Soft clustering techniques which are fuzzy based and rough set based have been proposed by a lot of researchers. They also handle high dimensional data effectively.

Categorical Data and High Dimensional Data

The high dimensionality of objects is represented using the feature vectors. Microarrays for gene sequencing and expression in the biological domain; epidemiological, ecological data, spatial data, spatio-temporal data in the Medical Information domain are some fields that involve complex and high dimensional data objects. Exhaustive ecological, epidemiological studies with respect to *sand flies* and control measures in a specific geographic region is carried out[8]. Rough Set Theory based modeling was derived based on *Myocardial Infarction* data obtained from one city and applied to predict similar condition in another city[9]. Clinical epidemiological study on demographic based risk factors for *Myocardial Infraction* is documented in[7] .

II. Soft Clustering algorithms

Analysis of ecological, epidemiological, cohort data involves identifying specific condition features that result in an outcome. It includes grouping similar data objects for which *Clustering* is used. There are two categories of Clustering Algorithms – Crisp and Soft. Crisp category places a data item in exactly one cluster whereas the latter gives every data item, a degree of belonging to more than one cluster. Soft Clustering algorithms identify more than one cluster for a data item. Fuzzy based algorithms accomplish this using a membership function. Rough set based algorithms group data items into lower and upper regions of clusters. Rough Set Theory based algorithms are best suited for Bio-Informatics as it is data oriented, can handle inconsistent data, data with any number of critical attributes[9]. A data item can belong to more than one upper region of multiple clusters. This ability of the algorithms gives an edge when compared to hard clustering techniques. A comparative study of hard and soft algorithms is discussed and hoe soft clustering is better than hard clustering is also demonstrated.

Fuzzy C Means Algorithm

FCM allows a data item to belong to more than one cluster[1]. It is based on the minimization of

an objective function involving the number of data items. Every data item is represented as an*d* dimensional vector. The degree of membership of the objects is given by a matrix. The membership function for the data items at the edge of a cluster is less that the data item that is close to the centroid. The centroid is obtained by the calculating the average of all data items weighted by their degree of membership to a cluster. It is an iterative algorithm and goes through multiple passes before termination. The results tend to vary for different membership functions and just do not rely on the nature of data unlike Rough Set based algorithms. Further, it requires considerable domain knowledge to have a good choice of membership function.

Rough K Means Algorithm

Rough Set concepts are used in K-Means algorithm with the inclusion of lower and upper approximations. Each cluster is identified based on its lower and upper approximations. The properties defined are

(1) A data item $\vec{x}$ can belong to the lower approximation of at most one cluster. ***A*** is the subset of attributes considered for partition of the universal set, c_i is a cluster.

$$\vec{x} \in \underline{A}(\vec{c_i}) \rightarrow \vec{x} \in \overline{A}(\vec{c_i})$$

(2) A data item that is not a part of any lower approximation belongs to upper approximation of two or more clusters.

Every data item is represented as a vector and a distance vector is assigned between the data item and the centroid of a specific cluster. The distance ratio is used to determine the membership of the distance vector. A threshold value is used to check the ratio of distances between the data objects and cluster centroids.

MMR Algorithm

MMR algorithm uses the concept of *roughness* which allows a data item to belong to different clusters with different degrees. Roughness is the ratio of cardinality of lower and upper approximation[4]. It is documented to be successfully applied in different fields[7,8] . The algorithm uses roughness of an attribute to calculate its *mean roughness* with respect to other attributes. Later the attribute with minimum of this roughness is chosen as the attributes with *minimum of mean roughness* is considered for clustering. The attribute with minimum roughness is considered to be the splitting point further partitioning till the desired number of partitions were achieved.

MMeR Algorithm

Rough Set Theory partitions a given set of objects (universe) represented in an Information System based on equivalence relation which is a subset of attributes. MMeR algorithm is based on the principle that mean of roughness values for every equivalence class of each attribute is to be calculated. Then the attribute that belongs to an equivalence class with least mean is chosen as splitting attribute for further partitioning. The lower the mean roughness is the higher the crispness of the cluster[5].

$$\mathrm{MeR}(a_i = a) = \frac{\sum_{j=1}^{n} R_{a_j}\left(\frac{x}{a_i = a}\right)}{n-1}$$

In order to calculate MeR, upper and lower approximations based on R (subset of attributes considered for partitioning the universe) are to be calculated. Then Roughness R_{a_j} is calculated based on the ratio of lower and upper approximation which is equivalent to the accuracy of approximation and is the measure of roughness[5].

Table 2
A sample dataset with four attributes (a1 – a4)

	History (a1)	Medical Condition1(a2)	Medical Condition2 (a3)	Medical Condition3 (a4)
1	diabetes	High	High	RS171
2	smoking	Medium Low	Moderate	RS10
3	hypertension	Medium High	Low	RS110
4	smoking	High	Moderate	RS110
5	hypertension	Medium High	Low	RS171
6	diabetes	Low	High	RS10
7	hypertension	Medium High	High	RS171
8	hypertension	Medium High	Low	RS171
9	diabetes	Low	High	RS10
10	smoking	Low	Moderate	RS10

Table 3
Calculation of Mean Roughness for a1

	X1 (h)	X2 (s)	X3 (c)	Mean Roughness
wrt a2	0	0.8	1	0.6
wrt a3	0.6	0	1	0.52
wrt a4	1	1	1	1

Table 3
Calculation of Mean of Mean Roughness

Attributes	Mean Roughness	Mean of Mean Roughness
a1	Rough a_i(a1),j=2 3 4(0.6, 0.5, 1)	0.7
a2	Rougha_j (a2), j = 1,3,4 (0.7500, 0.8929, 1)	0.88
a3	Rougha_j (a3), j = 1,2,4 (0.52, 0.94, 1)	0.82
a4	Rougha_j (a4), j = 1,2,3(1, 0.67, 1)	0.89

Cluster Selection in MMeR

Mean of minimum roughness of an attribute with respect to every other attribute is used to identify the splitting attribute and clusters. Hamming distance is used to calculate the average distance between every tuples in each cluster. Considering the average distances in each cluster, the one with least average distance is chosen for further splitting. The attribute a1 is chosen as the splitting attribute for the dataset given. This approach will create stable clusters as the mean of roughness is considered. This is better when compared to MMR algorithm in terms of the stability of the clusters.

SDR and SSDR Algorithm

SSDR algorithm is an improvement upon SDR algorithm, where after calculating the mean roughness of each attribute, standard deviation to each attribute is calculated[6]. This approach showed improvement in terms of purity index of the cluster. SSDR used the standard deviation of standard deviation roughness of each attribute. It is calculated using the formula

III. Purity of a Cluster

Purity of a cluster determines the relevance of the data items to its cluster[6].

Purity (i) = NC/ND

NC = Number of data occurring in both the i^{th} cluster and its equivalence class

ND =Number of data in the data set

Overall Purity = SP / N_Cl

SP = Sum of purity of all the clusters
N_Cl = Number of clusters

CONCLUSION

Rough set based algorithms handle categorical, hybrid and high dimensional data clustering in Bio-Informatics and other domains effectively. The algorithms also handle the instability of clustering processes and do not demand the domain knowledge. These features make them robust. Fuzzy-rough based concepts may be used for splitting phase.

REFERENCES

1. Manish Joshi. Correlating Fuzzy and Rough Clustering, Fundamenta Informaticae, pages 233 – 246, (2012)
2. John Stell and Michael Worboys, A Theory of Change for Attributed Spatial Entities. Geographic Information Science Lecture Notes in Computer Science, 5266:306 -319, (2008)
3. Pawan Lingras., Manish Joshi. Experimental Comparison of Iterative Versus Evolutionary Crisp and Rough Clustering.4(1), (2011)
4. Darshit Parmar et. al.,MMR: An algorithm for clustering categorical data using Rough Set Theor.y Data & Knowledge Engineering, 63:879–893, (2007).
5. Tripathy B.K., Prakash Kumar. MMeR: An algorithm for clustering categorical data using Rough Set theory. International Journal of Rapid Manufacturing, 1(2): 189-207, (2007)
6. Tripathy B.K., Adhir Ghosh. SSDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory, Advances in Applied Science Research, 2 (3): 314-326, (2011)
7. Srivatsava, R.K. et.al., Comparison of Acute Myocardial Infarction Risk Factors in Young and Elderly Patients – A Clinco-Epidemiological Study, International Journal of Pharma and Bio Sciences, 6(3):479 – 485, (2015).
8. Sathe, T.V., Ecology, Epidemiology and Control of Sand Flies from Kolhapur Region, India, International Journal of Pharma and Bio Sciences, 2014, 5(4): 1037 – 1045, (2014)
9. Vinterbo, S. Predictive Models in Medicine: Some Methods for Construction and Adaptation, PhD Thesis, Norwegian University of Science and Technology, Department of Computer and Information Science, (1999).