A Rough Clustering Algorithm for Mining Outliers in Categorical Data

N.N.R. Ranga Suri[1], Musti Narasimha Murty[2], and Gopalasamy Athithan[1,3]

[1] Centre for AI and Robotics (CAIR), Bangalore, India
{rangasuri,athithan.g}@gmail.com
[2] Dept of CSA, Indian Institute of Science (IISc), Bangalore, India
mnm@csa.iisc.ernet.in
[3] Presently Working at Scientific Analysis Group (SAG), Delhi, India

Abstract. Outlier detection is an important data mining task with applications in various domains. Mining of outliers in data has to deal with uncertainty regarding the membership of such outlier objects to one of the normal groups (classes) of objects. In this context, a soft computing approach based on rough sets happens to be a better choice to handle such mining tasks. Motivated by this requirement, a novel rough clustering algorithm is proposed here by modifying the basic k-modes algorithm to incorporate the lower and upper approximation properties of rough sets. The proposed algorithm includes the necessary computational steps required for determining the object assignment to various clusters and the modified centroid (mode) computation on categorical data. An experimental evaluation of the proposed rough k-modes algorithm is also presented here to demonstrate its performance in detecting outliers using various benchmark categorical data sets.

Keywords: Data mining, Soft computing, Rough sets, Outlier detection, Data Clustering.

1 Introduction

Rough set theory [7] was basically developed to deal with vagueness in the data. While fuzzy sets deal with such data using a partial membership function, rough sets express the same by the boundary region of a set. A rough set (C) is a set of objects which cannot be with certainty classified as members of the set or its complement using the available knowledge. Thus, associated with every rough set, there is a pair of precise sets known as *lower approximation* ($\underline{C}$) and *upper approximation* ($\overline{C}$) of the rough set. The basic idea is to separate discernible objects from indiscernible ones and to assign them to lower and upper approximations of the set respectively.

As brought out in [1,9], outlier detection is a non-trivial data mining task involving various research issues such as the method of detection, nature of the data, etc. The uncertainty prevailing in crisp labeling of a data object as a normal one or an outlier needs to be handled by employing a soft computing tool

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 170–175, 2013.

such as rough sets. If the underlying data happens to be categorical in nature, as is the case with many real life data mining applications, it leads to more serious concerns due to the non-availability of an effective similarity/dissimilarity measure defined on such data [10].

Motivated by the above discussion, a novel rough clustering algorithm suitable for mining outliers in categorical data is proposed here by modifying the basic k-modes algorithm [4] due to its efficiency as a member of the k-means family of algorithms. The novel algorithm incorporates the lower and upper approximation properties of rough sets similar to the rough k-means algorithm [5].

Section 2 provides an overview of rough clustering including the rough k-means algorithm. The novel rough k-modes algorithm is presented in the subsequent section along with the necessary computational steps. An experimental evaluation of the proposed method for outlier detection is furnished in Section 4. Section 5 concludes this paper with some discussion and ideas for future work.

2 Related Work

In rough clustering, the upper and lower approximations of a cluster are required to satisfy the following basic properties of rough sets [8]:

- A data object can be a member of one lower approximation at most.
- A data object that is a member of the lower approximation of a cluster is also member of the upper approximation of the same cluster.
- A data object that does not belong to any lower approximation is a member of at least two upper approximations.

In contrast to the original rough set theory developed based on the classic set theory, rough clustering is inspired by intervals. The rough k-means algorithm proposed in [5] is an effort in this direction by suitably modifying two important aspects of k-means clustering: (a) calculation of the centroids and (b) object assignment to the lower/upper approximations of a cluster. An initial version of this algorithm was improved upon by proposing various refinements as given in [8]. According to the improved version of the rough k-means algorithm [5], the two fundamental steps mentioned above are computed as follows:

- *Step (A)*: Modified procedure for calculating the centroid Z_j of a cluster C_j.
  - If ($\underline{C_j} \neq \emptyset$ and $\overline{C_j} - \underline{C_j} = \emptyset$) then $z_{j,r} = \frac{\sum_{X_i \in C_j} x_{i,r}}{|\overline{C_j}|}$.
  - If ($\underline{C_j} = \emptyset$ and $\overline{C_j} - \underline{C_j} \neq \emptyset$) then $z_{j,r} = \frac{\sum_{X_i \in (\overline{C_j} - \underline{C_j})} x_{i,r}}{|\overline{C_j} - \underline{C_j}|}$.
  - else $z_{j,r} = w_{low} \frac{\sum_{X_i \in \underline{C_j}} x_{i,r}}{|\underline{C_j}|} + w_{up} \frac{\sum_{X_i \in (\overline{C_j} - \underline{C_j})} x_{i,r}}{|\overline{C_j} - \underline{C_j}|}$.

  where $1 \leq r \leq m$ with m being the dimensionality, and $w_{low} + w_{up} = 1$.
- *Step(B)*: Modified procedure for assigning an object X_i to clusters.
  Let $d(Z_j, X_i)$ be the distance between object X_i and the centroid Z_j of C_j.
  1. Determine the nearest centroid Z_j, s.t. $d(Z_j, X_i) = min_{1 \leq j \leq k} d(Z_j, X_i)$.

2. Determine the centroids Z_l's that are also close to X_i.
   Let $T = \{l : d(Z_l, X_i)/d(Z_j, X_i) \leq \epsilon \text{ and } l \neq j\}$. Then,
   - If $T \neq \emptyset$ then $[X_i \in \overline{C_j} \text{ and } X_i \in \overline{C_l}, \forall l \in T]$.
   - else $[X_i \in \overline{C_j} \text{ and } X_i \in \underline{C_j}]$.

Here, $\epsilon \geq 1$ is a roughness parameter taking a user specified value.

3 Proposed Algorithm

Let $D = \{X_1, X_2, X_3, \ldots, X_n\}$ be the input data set consisting of n data objects, described using m categorical attributes. Each data object X_i is represented as $[x_{i,1}, x_{i,2}, \ldots, x_{i,m}]$. Consider a clustering scenario with k clusters $\{C_1, C_2, \ldots, C_k\}$ represented by the centroids $\{Z_1, Z_2, \ldots, Z_k\}$ respectively.

For the purpose of modifying the basic k-modes algorithm [4] to include the rough set properties, it requires certain measures defined on categorical data to formulate the two important clustering steps mentioned in Section 2. More specifically, a suitable categorical dissimilarity measure between an object X_i and a cluster C_j (with its mode Z_j) is required for assigning the data objects to the clusters. Similarly, computation of new cluster centroids (modes) also needs to be modified by incorporating the frequency counts of various categorical attribute values of the objects assigned to that cluster. In this regard, the following categorical dissimilarity measure as defined in [6] is considered here.

$$d(Z_j, X_i) = \sum_{r=1}^{m} \phi(z_{j,r}, x_{i,r}) \tag{1}$$

where

$$\phi(z_{j,r}, x_{i,r}) = \begin{cases} 1, & \text{if } z_{j,r} \neq x_{i,r}, \\ 1 - \frac{|C_{j,r}|}{|C_j|}, & \text{otherwise.} \end{cases}$$

where $|C_{j,r}|$ is the number of objects with category value $x_{i,r}$ for the r^{th} attribute in the j^{th} cluster.

Let $freq(x_{i,r})$ denote the number of objects in D with the value $x_{i,r}$ for the r^{th} attribute. Then, the density of a data object X_i can be computed as

$$density(X_i) = \frac{1}{mn} \sum_{r=1}^{m} freq(x_{i,r}) \tag{2}$$

Similarly, let $freq_j^{low}(x_{i,r})$ and $freq_j^{up}(x_{i,r})$ denote the number of objects in the lower and upper approximations of j^{th} cluster respectively with the value $x_{i,r}$ for the r^{th} attribute.

Utilizing these categorical measures, various computational steps of the proposed rough k-modes method are formulated as given in Algorithm 1, in accordance with rough set properties listed in Section 2. To ensure faster convergence, the cluster initialization method described in [3] has been considered.

After generating the rough clusters, any clustering-based outlier detection method such as the one in [10], can be employed for producing the outliers.

Algorithm 1. Proposed rough k-modes algorithm

Input: A categorical data set D consisting of n objects of m dimensionality.
Output: k rough clusters (lower and upper approximations) with their modes.

1: Compute $density(X_i)$ of each data object $X_i \in D$ using Equation 2.
2: Determine the initial set of k cluster representatives as per the method in [3].
3: Assign data objects to lower and/or upper approximation of clusters.

- Compute the cluster-to-object distance $d(Z_l, X_i)$ as per Equation 1.
- Determine object assignment according to Step(B) in Section 2.

4: Count cluster-wise attribute value frequencies $freq_j^{low}(x_{i,r})$ and $freq_j^{up}(x_{i,r})$.
5: Compute the new mode $Z_j^* \leftarrow X_i$ of each cluster C_j, s.t. $max_{X_i \in C_j} density_j(X_i)$, where $density_j(X_i)$ is the density of data object X_i w.r.t. cluster C_j given by

- if ($\underline{C_j} \neq \emptyset$ *and* $(\overline{C_j} - \underline{C_j}) = \emptyset$) then $density_j(X_i) = \frac{1}{m} \sum_{r=1}^{m} (\frac{freq_j^{low}(x_{i,r})}{|\underline{C_j}|})$.
- if ($\underline{C_j} = \emptyset$ *and* $\overline{C_j} \neq \emptyset$) then $density_j(X_i) = \frac{1}{m} \sum_{r=1}^{m} (\frac{freq_j^{up}(x_{i,r})}{|\overline{C_j}|})$.
- else $density_j(X_i) = \frac{1}{m} \sum_{r=1}^{m} (w_{low} \frac{freq_j^{low}(x_{i,r})}{|\underline{C_j}|} + w_{up} \frac{freq_j^{up}(x_{i,r}) - freq_j^{low}(x_{i,r})}{|\overline{C_j} - \underline{C_j}|})$. where $w_{low} + w_{up} = 1$.

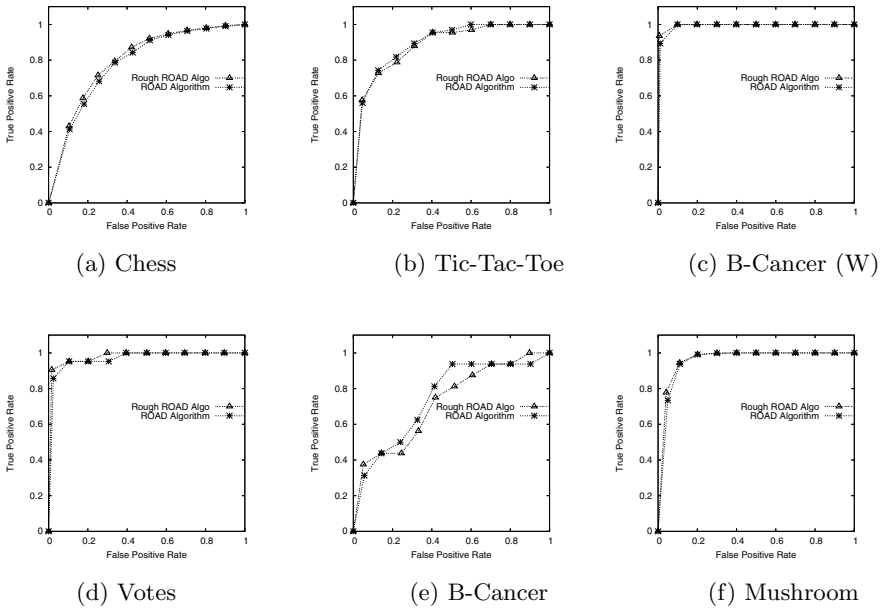
6: Assign data objects to lower and/or upper approximation of clusters as in Step 3.
7: Check for convergence of the algorithm. If not, repeat Steps 4-6 till convergence.

4 Experimental Evaluation

For the purpose of evaluating the effectiveness of the proposed rough k-modes algorithm for outlier detection, a recent related work named as Ranking-based Outlier Analysis and Detection (ROAD) framework [10] has been considered. This framework basically employs two independent ranking schemes to produce a likely set of outliers from the input data. Instead of using the basic k-modes algorithm, the current evaluation employs the novel rough k-modes algorithm with the ROAD framework, hereafter referred to as 'Rough ROAD', for determining the clustering-based ranking of the data objects. The rest of the details regarding this framework remain the same as in [10].

Table 1. Performance comparison on various benchmark data sets

Name of Data Set	Dim	Outlier Class Label	# Normal Objects	# Outlier Objects	# Outliers Detected	
					Rough ROAD Algorithm	ROAD Algo
Chess (End-Game)	36	nowin	1669	305	**131**	126
Tic-Tac-Toe	9	negative	626	66	**38**	37
Breast Cancer (W)	10	4(malignant)	444	47	**44**	42
Congressional Votes	16	republican	124	21	**19**	18
Breast Cancer	9	recurrence-events	196	16	**6**	5
Mushroom	22	poisonous	4208	783	**609**	575



(a) Chess (b) Tic-Tac-Toe (c) B-Cancer (W)

(d) Votes (e) B-Cancer (f) Mushroom

Fig. 1. Comparative view of the performance on various benchmark data sets

This experimental evaluation has been carried out employing six frequently used categorical data sets taken from the UCI ML Repository [2]. A data preparation procedure similar to the one followed in [10] has been applied resulting in the processed data sets as per the details furnished in Table 1.

The evaluation procedure as explained in the ROAD framework has been considered to evaluate the performance of the proposed algorithm. Accordingly, the experimental results obtained using the ROAD methodology with and without the rough k-modes algorithm have been furnished in Table 1. The corresponding ROC curves generated on the benchmark data sets are shown in Fig. 1 for a quick comparison. This evaluation indicates that the proposed rough k-modes clustering algorithm has indeed produced notably better results on all the data sets, highlighting its utility for the outlier detection task.

One of the important parameters of the proposed rough clustering algorithm is ϵ (introduced in Step(B) in Section 2). Varying the value of this parameter has its impact on the rough characteristics of the algorithm, there by on the overall outlier detection performance. To illustrate this impact, further experimentation has been carried out on Mushroom data set [2]. Fig. 2 shows the number of outliers detected on this data set with varying value of this parameter. The other roughness parameters of the algorithm have been set to fixed values as suggested in [8] ($w_{low} = 0.7$, $w_{up} = 0.3$) through out this experimentation and the parameter specific to the ROAD framework as $\alpha = 2$.

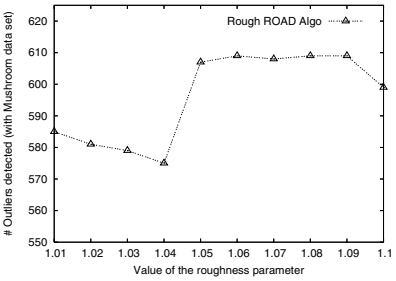


Fig. 2. Outlier detection performance with varying roughness parameter value

5 Conclusion and Future Work

Motivated by the frequent use of rough sets based soft computing approaches in various data mining applications, a novel rough k-modes algorithm has been proposed here for outlier detection. The performance of the proposed algorithm has been experimentally evaluated on various benchmark categorical data sets employing the ROAD framework and also compared it with that of the framework using the basic k-modes algorithm. However, further enhancement to detection accuracy may be possible by experimenting with other roughness parameters.

References

1. Albanese, A., Pal, S.K., Petrosino, A.: Rough sets, kernel set and spatio-temporal outlier detection. IEEE Trans. on Knowledge and Data Engineering (2012) (online)
2. Asuncion, A., Newman, D.J.: UCI machine learning repository (2007), `http://archive.ics.uci.edu/ml`
3. Cao, F., Liang, J., Bai, L.: A new initialization method for categorical data clustering. Expert Systems with Applications 36, 10223–10228 (2009)
4. Huang, Z.: A fast clustering algorithm to cluster very large categorical data sets in data mining. In: SIGMOD DMKD Workshop, pp. 1–8 (1997)
5. Lingras, P., Peters, G.: Applying rough set concepts to clustering. In: Rough Sets: Selected Methods and Applications in Management and Engineering, pp. 23–38. Springer, London (2012)
6. Ng, M.K., Li, M.J., Huang, J.Z., He, Z.: On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE PAMI 29(3), 503–507 (2007)
7. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
8. Peters, G.: Some refinements of rough k-means clustering. Pattern Recognition 39, 1481–1491 (2006)
9. Suri, N.N.R.R., Murty, M.N., Athithan, G.: Data mining techniques for outlier detection. In: Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications, ch. 2, pp. 22–38. IGI Global, New York (2011)
10. Suri, N.N.R.R., Murty, M.N., Athithan, G.: An algorithm for mining outliers in categorical data through ranking. In: IEEE HIS, Pune, India, pp. 247–252 (2012)