Brock University

Department of Computer Science

Rough set clustering

Ivo Düntsch and Günther Gediga

Technical Report # CS-15-02
January 2015

Brock University
Department of Computer Science
St. Catharines
Ontario Canada L2S 3A1

Universität Münster, Institut IV
Department of Psychology
Münster
Germany

www.cosc.brocku.ca

Rough set clustering*

Ivo Düntsch†‡
Brock University
St. Catharines
Ontario, Canada, L2S 3A1
duentsch@brocku.ca

Günther Gediga
Department of Psychology, Institut IV
Universität Münster, Fliednerstr. 21
Münster, Germany
gediga@uni-muenster.de

January 27, 2015

Abstract

We present a survey of clustering methods based on rough set data analysis.

1 Introduction to rough sets

The main application area of rough set data analysis is feature reduction and supervised learning. In most cases, clustering methods based on rough sets assume one or more given partitions of the data set and then aim to find a (cluster) variable which best represents the data according to some predefined measure.

Rough sets have been introduced in the early 1980s as a tool to handle uncertain information [32]. It is based on the idea that objects can only be distinguished up to the features which describe them. More formally, given an equivalence relation θ on a universe U, we assume that we know the world only up to the equivalence classes of θ and have no other knowledge about the objects within a class. A pair $\langle U, \theta\rangle$ is called an *approximation space*; the set of equivalence classes of θ is denoted by $\mathcal{P}(\theta)$. Given some $X \subseteq U$, the *lower approximation of* X is the set $\underline{X}_\theta = \{x \in U : \theta(x) \subseteq X\}$ and the *upper approximation of* X is the set $\overline{X}^\theta = \{x \in U : \theta(x) \cap X \neq \emptyset\}$; here, $\theta(x)$ is the class of θ containing x. For the sake of clarity we will sometimes write $\mathrm{low}_\theta(X)$ for $\underline{X}_\theta$ and $\mathrm{upp}_\theta(X)$ for $\overline{X}^\theta$. If θ is understood we will omit its index.

A *rough set* is a pair $\langle \underline{X}, \overline{X}\rangle$. We interpret the approximation operators as follows: If $\theta(x) \subseteq X$, then we know for certain that $x \in X$, if $\theta(x) \cap \overline{X} = \emptyset$ we are certain that $x \notin X$. In the *area of uncertainty* $\overline{X} \setminus \underline{X}$ we can make no certain prediction since $\theta(x)$ intersects both X and $U \setminus X$.

*To appear in "Handbook of Cluster Analysis", Eds. C. Hennig, M. Meila, F. Murtagh, R. Rocci

†The ordering of authors is alphabetical and equal authorship is applied.

‡Ivo Düntsch is grateful for support by the Natural Sciences and Engineering Reearch Council of Canada

The primary statistical tool of rough set data analysis (RSDA) is the *approximation quality function* $\gamma : 2^U \to [0,1]$: If $X \subseteq U$, then

$$\gamma(X) = \frac{|\underline{X}| + |U \setminus \overline{X}|}{|U|}, \tag{1.1}$$

which is just the ratio of the number of certainly classified elements of U to the number of all elements of U. The approximation quality γ is a manifestation of the underlying statistical principle of RSDA, namely, the principle of indifference: Within each equivalence class, the elements are assumed to be randomly distributed. If $\gamma(X) = 1$, then $X = \underline{X} = \overline{X}$; in this case, we say that X is *definable*. Another index frequently used in RSDA is the *accuracy measure* which is defined as

$$\alpha(X) = \frac{|\underline{X}|}{|\overline{X}|}. \tag{1.2}$$

The index $\alpha(X)$ is also called the *roughness of* X [28].

The main application area of RSDA is classification, i.e. supervised learning: An *information system* is a tuple $\mathcal{I} = \langle U, \Omega, (V_q)_{q\in\Omega}, (f_q)_{q\in\Omega}\rangle$, where U is a finite nonempty set of objects, Ω a finite nonempty set of attributes, for each $q \in \Omega$, V_q is a finite set of values which attribute q can take, and $f_q : U \to V_q$ is the information function which assign to each object x its value under attribute q. Each subset Q of Ω defines an equivalence relation θ_Q on U by setting

$$x \equiv_{\theta_Q} y \iff f_q(x) = f_q(y) \text{ for all } q \in Q. \tag{1.3}$$

The equivalence class of x with respect to the equivalence relation θ_Q will usually be denoted by $B_x(Q)$.

A *decision system* is an information system $\mathcal{I}$ enhanced by a decision attribute d with value set V_d and information function $f_d : U \to V_d$. To avoid trivialities, we will assume that f_d takes at least two values, i.e. that θ_d has at least two classes. If $Q \subseteq \Omega$ and X is a class of θ_Q we say that X is Q,d *– deterministic*, if there is a class Y of θ_d such that $X \subseteq Y$. The *approximation quality of* Q *with respect to* d is now defined as

$$\gamma(Q,d) = \frac{|\bigcup\{X : X \text{ is } Q,d - \text{deterministic}\}|}{|U|}. \tag{1.4}$$

$\gamma(Q,d)$ is the relative number of elements of U which can be correctly classified given the knowledge of Q. If $\theta_d = \theta_\Omega$, i.e. if the partition to be approximated is the partition given by the full set of attributes, we just write $\gamma(Q)$.

Another prominent feature of RSDA is the reduction of the number conditional attributes required to approximate the decision attribute: A *reduct of* $\mathcal{I}$ *with respect to* d is a set Q of attributes minimal with respect to the property

$$\gamma(Q,d) = \gamma(\Omega,d). \tag{1.5}$$

For a more complete treatment of rough sets we invite the reader to consult [9].

1.1 The variable precision model

It is assumed in the rough set model that the boundaries of the equivalence classes are crisp, and thus, that measurements are accurate. There are several possibilities to reduce the precision of prediction to cope with measurement error. One such approach is the *variable precision rough set model* [42, 43] which assumes that rules are valid only within a certain part of the population: Let U be a finite universe, $X, Y \subseteq U$, and first define the *relative degree of misclassification* of X with respect to Y

$$c(X,Y) = \begin{cases} 1 - \frac{|X \cap Y|}{|X|}, & \text{if } |X| \neq 0, \\ 0, & \text{if } |X| = 0. \end{cases} \tag{1.6}$$

Clearly, $c(X,Y) = 0$ if and only if $X = 0$ or $X \subseteq Y$, and $c(X,Y) = 1$ if and only if $X \neq \emptyset$ and $X \cap Y = \emptyset$. The *majority requirement* of the VP – model implies that more than 50% of the elements in X should be in Y; this can be specified by an additional parameter β which is interpreted as an admissible classification error, where $0 \leq \beta < 0.5$. The *majority inclusion relation* $\overset{\beta}{\subseteq}$ (with respect to β) is now defined as

$$X \overset{\beta}{\subseteq} Y \Longleftrightarrow c(X,Y) \leq \beta. \tag{1.7}$$

Given a family of nonempty subsets $\mathcal{X} = \{X_1, ..., X_k\}$ of U and $Y \subseteq U$, the *lower approximation* $\underline{Y}_\beta$ *of* Y *given* $\mathcal{X}$ *and* β is defined as the union of all those X_i, which are in relation $X_i \overset{\beta}{\subseteq} Y$, in other words,

$$\underline{Y}_\beta = \bigcup \{X \in \mathcal{X} : c(X,Y) \leq \beta\} \tag{1.8}$$

The classical approximation quality $\gamma(Q,d)$ is now replaced by a three-parametric version which includes the external parameter β, namely,

$$\gamma(Q,d,\beta) = \frac{|\mathrm{Pos}(Q,d,\beta)|}{|U|}, \tag{1.9}$$

where $\mathrm{Pos}(Q,d,\beta)$ is the union of those equivalence classes X of θ_Q for which $X \overset{\beta}{\subseteq} Y$ for some decision class Y. Note that $\gamma(Q,d,0) = \gamma(Q,d)$.

Unfortunately, $\gamma(Q,d,\beta)$ is not necessarily monotone [2, 10]. In [10] we show that $\gamma(Q,d,\beta)$ is a special instance of Goodman–Kruskal's λ [13] and how a class of monotone measures similar to $\gamma(Q,d,\beta)$ can be constructed.

The upper approximation within the variable precision model is defined in the following way:

$$\overline{Y}^\beta = \bigcup \{X \in \mathcal{X} : c(X,Y) < 1 - \beta\} \tag{1.10}$$

Following this definition it is obvious that the restriction $0 \leq \beta < 0.5$ must hold. Given this definition, the accuracy measure α of the classical RSDA can be rephrased as

$$\alpha(Y)_\beta = \frac{|\underline{Y}_\beta|}{|\overline{Y}^\beta|} \tag{1.11}$$

1.2 Rough Entropy and Shannon Entropy

In many situations Shannon Entropy may be used as a measure of the success of optimal guessing within data, however, we have shown in [8] that this approach does not fit the theoretical framework of rough set analysis. Using rough sets, the data are split into a deterministic part (lower approximation) and an indeterministic part (upper approximation). Here, the deterministic part corresponds to "knowing" whereas the indeterministic part still can be applied by "guessing". Throughout this section we let A be an attribute set, and θ_A be the equivalence relation in U based on A with associated partition $\mathcal{P}(A)$.

As described in section 3.2 below, Jiang et al. [16] exhibit a revised version of entropy for the purpose of outlier detection using the rough set model.

Liang et al. [20] generalize the concept of rough entropy of Liang & Shi [19] to incomplete information systems, which is simply

$$RE(A) = -\sum_{X\in\mathcal{P}(A)} \frac{|X|}{|U|}\cdot \log_2(|X|). \tag{1.12}$$

Quian & Liang [35] offer an alternative look on entropy, which they call combination entropy, defined by

$$CE(A) = \sum_{X\in\mathcal{P}(A)} \frac{|X|}{|U|}\cdot\left(1-\frac{|X|\cdot(|X|-1)}{|U|\cdot(|U|-1)}\right). \tag{1.13}$$

Similar to Shannon Entropy, CE increases as the equivalence classes become smaller through finer partitioning.

Liang et al. [21] offer a further alternative to the accuracy $\alpha(X)$ of (1.2). The accuracy $\alpha(X)$ does not take into account what the granulation of the approximating sets look like. Therefore, an index proposed by [21] is based on knowledge granulation of an attribute set A in the set U, defined by

$$KG(A) = \sum_{X\in\mathcal{P}(A)} \frac{|X|^2}{|U|^2}. \tag{1.14}$$

Then, the roughness of the approximation of X is a weighted α measure given by

$$Roughness(X) = (1-\alpha(X))\cdot KG(A) \tag{1.15}$$

which results in a new accuracy measure $\alpha'(X)$:

$$\alpha'(X) = 1 - Roughness(X) = 1-(1-\alpha(X))\cdot KG(A) = \alpha(X)\cdot KG(A). \tag{1.16}$$

Any of the proposed measures in this section can be used within a reduct search method which may be considered as a tool to find clusters as well. We will discuss this further in Section 2.2.

2 Methods of rough set clustering

The starting point of rough set clustering is an information system $\mathcal{I}$ with the aim of finding one attribute which best represents all attributes according to some predefined criterion.

2.1 Total roughness and min-min roughness based methods

Suppose that $\mathcal{I} = \langle U, \Omega, (V_q)_{q\in\Omega}, (f_q)_{q\in\Omega}\rangle$ is a given information system. If $p, q \in \Omega$, $p \neq q$, and $V_p = \{x_1, \ldots, x_n\}$, the *roughness of x_k with respect to q* is defined as

$$R(x_k, p, q) = \frac{\mathrm{low}_{\theta_q}(f_p^{-1}(x_k))}{\mathrm{upp}_{\theta_q}(f_p^{-1}(x_k))}, \tag{2.1}$$

and the *mean roughness of p with respect to q* is the value

$$MR(p, q) = \frac{\sum_{k=1}^{n} R(x_k, p, q)}{n}, \tag{2.2}$$

The *total roughness of attribute p* [28] is defined as

$$TR(p) = \frac{\sum_{q\in\Omega, r\neq p} MR(p, q)}{|\Omega| - 1}. \tag{2.3}$$

It is suggested in [28] to use the attribute with the highest total roughness as a clustering attribute. A technique somewhat opposite to TR was proposed by Parmar et al [31] which purports to "handle uncertainty in the process of clustering categorical data." Define

$$mR(p, q) = 1 - \frac{\sum_{k=1}^{n} R(x_k, p, q)}{n}, \tag{2.4}$$

and

$$\min R(p) = \min\{mR(p, q) : q \in \Omega, p \neq q\}, \tag{2.5}$$

as well as

$$\min\min R = \min\{MR(p) : p \in \Omega\}. \tag{2.6}$$

see [31] for details.

2.2 Reduct based clustering

A straight forward application of RSDA for cluster analysis are techniques related to the reducts of an information system as defined in (1.5), which are mainly used for nominally scaled data sets. The method is simple: Let the classes of the partition induced by all attributes be the sets to be approximated, and use a subset of the given attributes to approximate these classes. A minimal subset of the attributes which approximate the classes in an optimal way given a criterion C is called a *C-reduct*.

The partition which can be constructed by the attributes of the reduct can be interpreted as clusters. Note, that there may be several C-reducts, which may generate different partitions as well.

Given the criterion C $\stackrel{\text{def}}{=}$ "$\gamma(Q) = 1$", a reduct needs to approximate the full partition θ_Ω without any error, in other words, any C-reduct will generate the full partition. This is a good way to preprocess the data for a subsequent – conventional – cluster analysis, as any reduct gives the same information (in terms of Shannon-entropy) as the full set of attributes. In Questier et al. [34] there is an application of this approach; they show in an application that a set of 126 features can be reduced to 68 features using reducts of the RSDA.

Although reducing attributes is a nice feature on its own, reducts can be used for clustering as well: When using a $0 < \gamma^*(Q) < 1$ in the criterion C $\stackrel{\text{def}}{=}$ "$\gamma(Q) = \gamma^*(Q)$", we may obtain reducts which produce partitions with a number of subsets that is smaller or equal to the number of sets in the full partition. In the lower limit ($\gamma^*(Q) = 0$), no attribute is needed for approximation and therefore the only cluster remaining is U. The reduct search given a fixed $0 < \gamma^*(Q) < 1$ will generally result in different reducts – and in most of the cases – in different partitions as well. In this way we observe different sets of clusters, which can be used to approximate the full data set with the minimal precision $\gamma^*(Q)$. Scanning all reducts, a computation of fuzzy-membership of any element to a cluster is possible by aggregation using the set of partitions generated by the reducts.

Note, that the criterion of a reduced γ is not the only base for a reduct search. Chen & Weng [4] applied reduct search based on Shannon-Entropy, and Düntsch & Gediga [8] used rough entropy measures for their reduct search method. A further application has been proposed by Mayszko & Stepaniuk [27], who used Renyi-Rough-Entropy for searching an optimal partition of images.

Related approaches are the "maximum dependency of attributes" method of [14], and the "upper approximation based clustering" of [18].

2.3 Application of variable precision rough sets

Yanto et al. [39, 40] use the variable precision model to determine an attribute which can be used the find the best cluster representation among a set of nominally scaled attributes. Suppose that $\Omega = \{a_1, \ldots, a_r\}$ is the set of attributes and that $V_i = \{v_1^i, \ldots, v_{n(i)}^i\}$ is the set of attribute values of a_i. To avoid notational cluttering, with some abuse of language we denote the attribute function also by a_i. The equivalence relation belonging to a_i is denoted by θ_i; note that θ_i has exactly n_i classes, say, $X_1^i, \ldots, X_{n(i)}^i$. Choose some error tolerance parameter $\beta < 0.5$, and for each $1 \le j \le r$, $i \ne j$, $1 \le k \le n(i)$ set

$$\alpha_\beta(a_j, v_k^i) = |\underline{X_k^i}_{\theta_j,\beta}| \,/\, |\overline{X_k^i}^{\theta_j,\beta}|. \tag{2.7}$$

The *mean accuracy of a_i with respect to a_j* is now obtained as

$$\alpha_\beta(a_j, a_i) = \frac{\sum_{k=1}^{n(i)} \alpha_\beta(a_j, v_k^i)}{n(i)} \tag{2.8}$$

Finally the mean value aggregating over the attributes $a_j (j \ne i)$ results in the *mean variable precision roughness* of attribute a_i by:

$$\alpha_\beta(a_i) = \frac{\sum_{j \ne i} \alpha_\beta(a_j, a_i)}{r - 1}. \tag{2.9}$$

The attribute for clustering now is obtained by taking an a_i for which $\alpha_\beta(a_i)$ is maximal.

Yanto et al. [39] use their clustering algorithm with the data sets *Balloon*, *Tic-Tac-Toe Endgame*, *SPECT heart*, and *Hayes–Roth* from the UCI ML repository [12] and report that the result of the proposed VPRS technique provides a cluster purity of 83%, 69%, 64% and 63%, respectively. Another application of applying reduct search in the variable precision rough set model using student anxiety data is given in [40].

Reduct based methods can be used as well as a pre-processing tool for mixture or cluster analysis in case of a mutual high dependency of the variables. In [29] it is shown how rough set based algorithms for reduct detection can be used as a starter for mixture analysis. The analysis shows that using rough set based preprocessing results in a stable mixture estimation with smaller error and higher validity than using the full set of variables.

2.4 Rough K – means

We start with a set of n elements $x_1, ..., x_n$ and each element x is described by an m dimensional vector $\mathbf{v_x}$ of real valued (interval scaled) measurements. The classical K-Means algorithm consists of the following iteration scheme:

1. Start with a random assignment of the n elements to cluster $C_0 = \{C_1, ..., C_k\}$. Set $t = 0$.
2. Compute values of the centroids of the k classes by
$$\bar{\mathbf{x}}_\mathbf{j} = \frac{\sum_{x \in C_j} \mathbf{v_x}}{|C_j|} (1 \leq j \leq k). \tag{2.10}$$
3. For $1 \leq i \leq n$, $1 \leq j \leq k$ compute the distances $d(\mathbf{v_{x_i}}, \bar{\mathbf{x}}_\mathbf{j}) = \|\mathbf{v_{x_i}} - \bar{\mathbf{x}}_\mathbf{j}\|$. Here, $\|.\|$ is usually the standard Euclidian norm.
4. Re-assign the n elements according to the minimal distance to the k cluster, resulting in C_{t+1}.
5. If $C_{t+1} = C_t$ stop. Otherwise set $t = t + 1$ and proceed with step 2.

The rough K – means method [26, 33, 24, 6] adopts the idea of a lower and an upper approximation of a set which is somewhat less rigid than using classes of an equivalence relation: With each subset X of U a pair $\langle \underline{X}, \overline{X} \rangle$ of subsets of U is associated such that

1. Each $x \in U$ is in at most one $\underline{X}$.
2. $\underline{X} \subseteq \overline{X}$ for all $X \subseteq U$.
3. An object $x \in U$ is not in any lower bound $\underline{X}$ if and only if there are $Y_0, Y_1 \subseteq U$ such that $\overline{Y_0} \neq \overline{Y_1}$ and $x \in \overline{Y_0} \cap \overline{Y_1}$.

Applying this idea produces a cluster structure consisting of pairs of lower and upper approximations of the cluster: $C^* = \{(\underline{C_1}, \overline{C_1}), ..., (\underline{C_k}, \overline{C_k})\}$.

Given a pair $(\underline{C_j}, \overline{C_j})$ the values of the centroids of cluster j has to be computed.

The rough K – means (sometimes called rough C – means, e.g. in [41]) method uses a parameterized mean by

$$\bar{\mathbf{x}}_\mathbf{j} = \begin{cases} (1-\omega)\frac{\sum_{x \in \underline{C_j}} \mathbf{v}}{|\underline{C_j}|} + \omega \frac{\sum_{x \in \overline{C_j} \setminus \underline{C_j}} \mathbf{v}}{|\overline{C_j} \setminus \underline{C_j}|}, & \text{if } \overline{C_j} \setminus \underline{C_j} \neq \emptyset \\ \frac{\sum_{x \in \underline{C_j}} \mathbf{v}}{|\underline{C_j}|}, & \text{otherwise.} \end{cases} \tag{2.11}$$

The parameter ω weights the influence of the upper approximation. If $\omega = 0$, the upper approximation will not be used for the centroid computation, and the iteration is identical to classical K – means method. Given new centroids, the re-assignment of the elements to (now) lower and upper bounds of clusters has to be described. In case of the lower bound, the standard rule from classical K–means is adopted. In order to find suitable upper approximations, a second parameter $\theta \geq 1$ is used. If $\mathbf{x}$ is assigned to $\underline{C_j}$, then $\mathbf{x}$ is assigned to any $\overline{C'_j}$ $(j' \neq j)$, if

$$\frac{d(\mathbf{v_i}, \bar{\mathbf{x}}'_\mathbf{j})}{d(\mathbf{v_i}, \bar{\mathbf{x}}_\mathbf{j})} \leq \theta. \tag{2.12}$$

If one assigns concrete values to the parameters (ω, θ) and randomly assigns each data object to exactly one lower approximation, we find a start with $C_0^* = \{(\underline{C_1}, \underline{C_1}), ..., (\underline{C_k}, \underline{C_k})\}$, and the scheme of the classical K-Means algorithm can be adopted for a rough K – means application.

[30] proposed an update of the parameters based on evolutionary optimization using the Davies-Bouldwin index [3] as a measure of fit, which is a simple representation of the odd of within-cluster variability and between cluster-variability. The parameters K and ω are under control of a genetic algorithm (GA; 10 bits in a chromosome, 20 chromosomes, crossover probability=0.8, mutation probability =0.02). The GA is governed by the Davies-Bouldwin index as optimization and convergence criterion. The algorithm was applied to several sets and compared to other clustering algorithms. The results show that the proposed method is rather successful – and very promising when using it with gene expression data. It should be noted, that [11] showed as well a successful treatment of gene expression data with rough set based clustering methods.

The paper of [7] deals with interval set clustering, which is a generalization of the Rough-K-Means methods. In this paper, the problem of outliers is treated. The method starts with the classical Rough-K-Means method. Using the upper and lower bounds of the sets, an index $LUF_h(x)$ is defined – which captures the degree to which object x is reachable from the h next neighbors. Fixing h and applying a threshold for LUF_h, it is possible to eliminate those objects from the lower approximation of a cluster which are too far away from their neighbors. These objects will be assigned to the upper approximation of the cluster. Using the new sets of lower and upper approximation, the centroids can be recalculated and the iteration will start again. Applications to synthetic data and the Wisconsin breast cancer data support the applicability of this method.

2.5 Tolerance Rough Set Clustering

Tolerance Rough Set Models [37] relax the transitivity requirement of equivalence relations, and are often used in information retrieval to find clusters of index terms in large text data bases [15]. A *tolerance space* is a pair $\langle U, R\rangle$, where U is a nonempty finite set, and R is a reflexive and symmetric relation. The *tolerance classes* are the sets of the form $R(x) = \{y \in U : xRy\}$. Unlike equivalence classes, different tolerance classes may have a nonempty intersection. Lower and upper approximation of some $X \subseteq U$ are defined as in the equivalence case.

As an example of tolerance rough set clustering we present the non–hierarchical document clustering proposed in [15]. We start with a set of $T = \{t_1, ..., t_N\}$ of index terms and a set $D = \{d_1, ..., d_M\}$ of documents each of which is indexed by a set T_d of terms from T. The number of times a term t_i occurs in document d_j is denoted by $f_{d_j}(t_i)$, and $\sum_{j=1}^{M}\sum_{i=1}^{N}$, the number of (indexed) terms occurring in document d, is denoted by T_d^*. The number of documents in which t_i occurs is denoted by $f_D(t_i)$, and the number of documents in which index terms t_i, t_j both occur is denoted by $n_D(t_i, t_j)$.

Given a threshold θ which is the minimal acceptable number of common terms in documents we now define a covering I_θ of T by

$$I_\theta(t_i) = \{t_i\} \cup \{t_j | n_D(t_i, t_j) \geq \theta\}. \tag{2.13}$$

Clearly, the sets $I_\theta(t_i)$ are the tolerance classes of a tolerance relation $\mathcal{I}_\theta$ on T defined by

$$t_i \mathcal{I}_\theta t_j \Longleftrightarrow t_j \in I_\theta(t_i). \tag{2.14}$$

Given a set of terms $X \subseteq T$ we are now able to define a lower and upper bound of X with respect to $\mathcal{I}_\theta$:

$$\underline{X}_{\mathcal{I}_\theta} = \{t_i | I_\theta(t_i) \subseteq X\} \tag{2.15}$$

$$\overline{X}^{\mathcal{I}_\theta} = \{t_i | X \cap I_\theta(t_i) \neq \emptyset\} \tag{2.16}$$

With each term t_i and each document d_j a weight $w(i, j)$ is associated by

$$w(i,j) = \begin{cases} 1 + \log(f_{d_j}(t_i)), & \text{if } t_i \in d_j, \\ 0, & \text{otherwise.} \end{cases} \tag{2.17}$$

If we have a set of (intermediate) representatives $C_t = \{X_1, ..., X_k\}$, we are now able to calculate $C_t^* = \{\underline{X_1}_{\mathcal{I}_\theta}, \overline{X_1}^{\mathcal{I}_\theta}, \ldots, \underline{X_k}_{\mathcal{I}_\theta}, \overline{X_k}^{\mathcal{I}_\theta}\}$ and use this information to find better representatives C_{t+1}.

Various measures of similarity between two documents may be used. One such index is the *binary weight Dice coefficient* is defined as

$$S(d_i, d_j) = \frac{2 \cdot C}{T_{d_i}^* + T_{d_j}^*}, \tag{2.18}$$

where C is the number of index terms which d_i and d_j have in common.

Suppose we are given a set of clusters $C = \{C_1, \ldots, C_k\}$ of documents. The algorithm constructs a representative $R_i \subseteq T$ for each $1 \leq i \leq k$ such that

1. Each document $d \in C_i$ has an index term in R_i, i.e. $T_d \cap R_i \neq \emptyset$.
2. Each term in R_i is possessed by a large number of documents in C_i (given by some threshold).
3. No term in R_i is possessed by every document in C_i.

3 Special topics

3.1 Validation based on rough sets

Using validation methods which does not take into account the special character of clusters build by rough set methods might be biased. Note, for example, that in the result of the rough-k-means method an object may belong to more than one cluster. Moreover, each cluster C_j is represented by its lower approximation $\text{low}(C_j)$ and an upper approximation $\text{upp}(C_j)$ and/or the boundary set $\text{upp}(C_j) \setminus \text{low}(C_j)$.

Lingras et al. [24] start with an "action" function $b_j(x_l)$, which assigns an element x_l to a set of clusters from a set $B_j \subseteq 2^U$. A simple loss function can been assigned by

$$Loss(b_j(x_l)|C_i) = \begin{cases} 0, & \text{if } C_i \in B_j, \\ 1, & \text{otherwise.} \end{cases} \tag{3.1}$$

Let $sim(x_l, C_i)$ be a similarity function of element x_l and cluster $\text{low}(C_i)$, e.g. the inverse of the distance of x_l and the centroid of $\text{low}(C_i)$. Assuming that the probability $p(C_i|x_l)$ is proportional to $sim(x_l, C_i)$ by

$$p(C_i|x_l) = \frac{sim(x_l, C_i)}{\sum_j sim(x_l, C_j)}, \tag{3.2}$$

we result in the risk function for assigning element x_l to a set B_j of clusters by

$$R(b_j(x_l)|x_l) = \sum_i Loss(b_j(x_l)|C_i) p(C_i|x_l). \tag{3.3}$$

As any element x_l gets its own risk evaluation, we are ready to define risks for certain subsets: The risk for lower approximation for the decision $b_j(x_l)$ for any element of the lower approximation is given by

$$\sum_{x_l \in \text{low}(C_i)} R(b_j(x_l)|x_l). \tag{3.4}$$

Furthermore the risk functions for elements of the upper approximation and the boundary are given by

$$\sum_{x_l \in \text{upp}(C_i)} R(b_j(x_l)|x_l) \tag{3.5}$$

and

$$\sum_{x_l \in \mathrm{upp}(C_i) \setminus \mathrm{low}(C_i)} R(b_j(x_l)|x_l), \tag{3.6}$$

respectively.

Applying the ideas to the synthetic and the Wisconsin breast cancer data, Lingras et al. [24] showed the rough and crisp cluster analysis exhibit similar results on the lower approximation, but differ in the risk of the assignment of boundary elements.

3.2 Outlier detection

In [16] it was shown that detecting outliers using the rough set model is at least as powerful as classical distances based methods [17] or KNN based methods [36].

Jiang et al. [16] start with the idea of rough entropy presented in [8]. Let θ be an equivalence relation on U with associated partition $\mathcal{P}(\theta)$. Starting with an element x and its equivalence class B_x, they define a leaving-out entropy of θ_x by

$$E(\theta \setminus \{B_x\}) = -\sum_{B \in \mathcal{P}(\theta), B \neq B_x} \frac{|B|}{|U| - |B_x|} \cdot \log_2\left(\frac{|B|}{|U| - |B_x|}\right). \tag{3.7}$$

Using the information $E(\theta)$ as a benchmark, it is straightforward to define the relative entropy of the class of object x given θ by

$$RE(B_x|\theta) = \begin{cases} 1 - \frac{E(\theta \setminus \{B_x\})}{E(\theta)}, & \text{if } E(\theta) > E(\theta \setminus \{B_x\}), \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

[16] introduce the relative cardinality of class B_x in θ by the difference of cardinality of B_x and the mean cardinality of the other classes:

$$RC(B_x|\theta) = \begin{cases} |B_x| - \frac{1}{|\mathcal{P}(\theta)|-1} \cdot \sum_{B \in \mathcal{P}(\theta), B \neq B_x} |B|, & \text{if } |\mathcal{P}(\theta)| > 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.9}$$

Now we are ready to define the outlier degree of B_x given θ by

$$OD(B_x|\theta) = \begin{cases} RE(B_x|\theta) \cdot \sqrt{\frac{|U| - |RC(B_x|\theta)|}{2|U|}}, & \text{if } RC(B_x|\theta) > 0, \\ RE(B_x|\theta) \cdot \sqrt{\frac{|U| - |RC(B_x|\theta)|}{2|U|}}, & \text{otherwise.} \end{cases} \tag{3.10}$$

In the latter, $RC(B_x|\theta) \leq 0$ which means that if x is assigned to a smaller equivalence class of θ, then the class B_x has a higher possibility to be an outlier class than the other classes. Furthermore, the higher the relative entropy of B_x, the greater is the likelihood of x to be an outlier.

Up to now the outlier definition is given for a fixed equivalence relation θ. As θ is normally obtained from a set of attributes, we may take the attributes into account. Let $A = \{a_1, ..., a_k\}$ be a set of attributes, $A_1 = A \setminus \{a_1\}$, and $A_j = A_{j-1} \setminus \{a_j\}$ for $1 < j < k$. Then, $A_{k-1} \subsetneq \ldots \subsetneq A_1 \subsetneq A$. Given any set of attribute Q and an element x, we denote the equivalence class of x with respect to θ_Q by $B_x(Q)$. The entropy outlier factor $EOF(x)$ is now defined by

$$EOF(x) = 1 - \frac{\sum_{j=1}^{k} \left(1 - OD(B_x(\{a_j\})\right) \cdot W_{\{a_j\}}(x) + \left(1 - OD(B_x(A_j)\right) \cdot W_{A_j}(x)}{2k} \tag{3.11}$$

using the weights

$$W_Q(x) = \sqrt{\frac{|B_x(Q)|}{|U|}}. \tag{3.12}$$

As $B_x(Q)$ changes with attributes in Q, $EOF(x)$ is a function of x and describes a distance of x to the rest of the elements of U in terms of an information function. Jiang et al. [16] show that applying $EOF(x)$ to the Lymphography data and the Wisconsin breast cancer data (which can be found in the UCI machine learning repository [12]) shows a better performance than distances based or KNN based methods.

4 Conclusion and outlook

We have presented an introduction to clustering based on Pawlak's rough set model and its associated data type, the information system.

As the rough set model is a basic idea of many papers, the overview we have given cannot be exhaustive. There exists a plethora of other approaches in the literature which share the idea of rough set analysis and adopt other concepts to analyze the data. Some of these approaches are "less rough and more other", for example,

- Rough-fuzzy-clustering and shadowed sets [1, 41],
- Clustering by categorical similarity measures based on rough sets using a hierarchical clustering scheme [5],
- A fast heuristic Rough DB–scan procedure [38].

We have shown that the rough set model as an – originally – symbolic data analysis tool has been developed into a viable system for cluster analysis. There are two main ideas of the rough set model which are promising for applications: The first is the idea of an upper bound and a boundary of set approximation. Tolerance Rough Set Clustering or Rough-K-Means Clustering use this idea. Although the application show satisfactory results, the methods depends on several parameters and thresholds, which may be somewhat problematic. The are some attempts to use these parameter as free parameters as well, but optimization can

only be done by genetic algorithms or comparable methods. Here, further investigations are necessary. The other idea is to use reduct based methods, either as pre-processing tool or as a tool for cluster generation. These methods show a more direct connection to the roots of rough set analysis, but the problem of time complexity has not been solved until now. As long as we are only interested in one reduct, the situation is not complicated, but finding all reducts within a variable precision model is still an NP-complete problem.

Apart from the complexity problem, it is worthy to note that applying ideas of the rough set model in cluster analysis seams to lead to stable and valid results in general. Furthermore, they offer new insights how concepts such as "cluster", "error" or "outlier" and even "validation" may be understood and defined. We think that this is a value on its own.

References

[1] S. Asharaf and M. N. Murty. An adaptive rough fuzzy single pass algorithm for clustering large data sets. *Pattern Recognition*, 36(12):3015, December 2003.

[2] M. Beynon. Reducts within the variable precision rough sets model: A further investigation. *European Journal of Operational Research*, 134:592–605, 2001.

[3] J. C. Bezdek and N. R. Pal. Some new indexes of cluster validity. *IEEE Transcations on Systems, Man, and Cybernetics, Part B: Cybernetics*, 28:301–315, 1998.

[4] Ch.-B. Chen and L.-Y. Wang. Rough set based clustering with refinement using Shannon's entropy theory. *Computers and Mathematics with Applications*, 52:1563–1576, 2006.

[5] D. Chen, D.-W. Cui, Ch.-X. Wang, and Zh.-R. Wang. A rough set-based hierarchical clustering algorithm for categorical data. *International Journal of Information Technology*, 12:149–159, 2006.

[6] J. Chen, W. Changsheng Zhang, and Zhu-Rong Wang. Efficient clustering method based on rough set and genetic algorithm. *Procedia Engineering*, 15:1498–1503, January 2011.

[7] M. Chen and D. Miao. Interval set clustering. *Expert Systems with Applications*, 38(4):2923–2932, April 2011.

[8] I. Düntsch and G. Gediga. Uncertainty measures of rough set prediction. *Artificial Intelligence*, 106:77–107, 1998.

[9] I. Düntsch and G. Gediga. *Rough set data analysis: A road to non-invasive knowledge discovery*. Methodos Publishers (UK), Bangor, 2000. `http://www.cosc.brocku.ca/~duentsch/archive/nida.pdf`.

[10] I. Düntsch and G. Gediga. Weighted λ precision models in rough set data analysis. In *Proceedings of the Federated Conference on Computer Science and Information Systems, Wrocław, Poland*, pages 309–316. IEEE, 2012.

[11] J. J. Emilyn and D. K. Rama. Rough set based clustering of gene expression data: A survey. *International Journal of Engineering Science and Technology*, 2:7160–7164, 2010.

[12] A. Frank and A. Asuncion. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2010.

[13] L. A. Goodman and W. H. Kruskal. Measures of association for cross classification. *Journal of the American Statistical Association*, 49:732–764, 1954.

[14] Tutut Herawan, Mustafa Mat Deris, and Jemal H. Abawajy. A rough set approach for selecting clustering attribute. *Knowledge-Based Systems*, 23(3):220–231, April 2010.

[15] Tu Bao Ho and Ngoc Binh Nguyen. Nonhierarchical document clustering based on a tolerance rough set model. *International Journal of Intelligent Systems*, 17(2):199–212, 2002.

[16] F. Jiang, Y. Sui, and C. Cao. An information entropy-based approach to outlier detection in rough sets. *Expert Systems with Applications*, 37:6338–6344, 2010.

[17] E. Knorr, R. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *VLDB Journal: Very Large Databases*, 8:237–253, 2000.

[18] Supriya Kumar De. A rough set theoretic approach to clustering. *Fundamenta Informaticae*, 62(3/4):409–417, October 2004.

[19] J. Liang and Z. Shi. The information entropy, rough entropy and knowledge granulation in rough set theory. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12:37–46, 2004.

[20] J. Liang, Z. Shi, D. Li, and M. J. Wierman. Information entropy, rough entropy and knowledge granulation in incomplete information systems. *International Journal of General Systems*, 6:641–654, 2006.

[21] J. Liang, J. Wang, and Y. Qian. A new measure of uncertainty based on knowledge granulation for rough sets. *Information Sciences*, 179:458–470, 2009.

[22] P. Lingras. Applications of rough set based k-means, Kohonen SOM, GA clustering. *Transactions on Rough Sets*, 7:120–139, 2007.

[23] P. Lingras, M. Chen, and D. Miao. Rough multi-category decision theoretic framework. In Guoyin Wang, Tian rui Li, Jerzy W. Grzymala-Busse, Duoqian Miao, Andrzej Skowron, and Yiyu Yao, editors, *RSKT*, volume 5009 of *Lecture Notes in Computer Science*, pages 676–683. Springer, 2008.

[24] P. Lingras, Min Chen, and A. Duoqian Miao. Rough cluster quality index based on decision theory. *IEEE Transactions on Knowledge & Data Engineering*, 21(7):1014–1026, July 2009.

[25] P. Lingras, M. Hogo, and M. Sn. Interval set clustering of web users using modified Kohonen self-organizing maps based on the properties of rough sets. *Web Intelligence and Agent Systems*, 2(3):217–225, 2004.

[26] P. Lingras and Ch. West. Interval set clustering of web users with rough K-Means. *Journal of Intelligent Information Systems*, 23(1):5–16, 2004.

[27] D. Mayszko and J. Stepaniuk. Adaptive multilevel rough entropy evolutionary thresholding. *Information Sciences*, 180(7):1138–1158, April 2010.

[28] L. J. Mazlack, A. He, and Y. Zhu. A rough set approach in choosing partitioning attributes. In *Proceedings of the ISCA 13th International Conference (CAINE-2000)*, pages 1–6, 2000.

[29] P. Mitra, S. K. Pal, and M. A. Siddiqi. Non-convex clustering using expectation maximization algorithm with rough set initialization. *Pattern Recognition Letters*, 24(6):863–873, March 2003.

[30] S. Mitra. An evolutionary rough partitive clustering. *Pattern Recognition Letters*, 25:1439–1449, 2004.

[31] D. Parmar, Teresa Wu, and M. Jennifer Blackhurst. MMR: an algorithm for clustering categorical data using rough set theory. *Data & Knowledge Engineering*, 63(3):879–893, December 2007.

[32] Z. Pawlak. Rough Sets. *Internat. J. Comput. Inform. Sci.*, 11:341–356, 1982.

[33] G. Peters. Some refinements of rough k-means clustering. *Pattern Recognition*, 39:1481–1491, 2006.

[34] F. Questier, I. Arnaut-Rollier, B. Walczak, and D.L. Massart. Application of rough set theory to feature selection for unsupervised clustering. *Chemometrics and Intelligent Laboratory Systems*, 63(2):155–167, 2002.

[35] Y. Quian and J. Liang. Combination entropy and combination granulation in rough set theory. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 16:179–193, 2008.

[36] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large datasets. In *Proceedings of the ACM SIGMOD conference on management of data*, pages 427–438, 2000.

[37] A. Skowron and J. Stepaniuk. Tolerance approximation spaces. *Fundamenta Informaticae*, 27:245–253, 1996.

[38] P. Viswanath and V. Suresh Babu. Rough-DBSCAN: a fast hybrid density based clustering method for large data sets. *Pattern Recognition Letters*, 30(16):1477–1488, December 2009.

[39] I. T. R. Yanto, T. Herawan, and M. M. Deris. Data clustering using variable precision rough set. *Intelligent Data Analysis*, 15:465–482, 2011.

[40] I. T. R. Yanto, P. Vitasari, T. Herawan, and M. M. Deris. Applying variable precision rough set model for clustering student suffering studyŠs anxiety. *Expert Systems with Applications*, 39:452–459, 2012.

[41] J. Zhou, W. Pedrycz, and D. Miao. Shadowed sets in the characterization of rough-fuzzy clustering. *Pattern Recognition*, 44(8):1738–1749, 2011.

[42] W. Ziarko. Variable precision rough set model. *Journal of Computer and System Sciences*, 46:39–59, 1993.

[43] W. Ziarko. Probabilistic approach to rough sets. *International Journal of Approximate Reasoning*, 49(2):272–284, October 2008.