A Rough Set Approach to Clustering

Staal Vinterbo Aleksander Øhrn

Knowledge Systems Group
Department of Computer and Information Science
Norwegian University of Science and Technology
N-7034 Trondheim, Norway
{staalv,aleks}@idt.ntnu.no

Abstract

A method of valuating attributes is proposed and is used to construct a semi-metric. This semi-metric is the basis of several proposed metrics that can be used in an unsupervised clustering algorithm for feature spaces where some or all dimensions do not have a metric defined on them. Additionally, the metrics can be used to construct a kNN-classifier that copes with incomplete or extraneous data.

1 Introduction

The technique of partitioning an unlabeled dataset into "interesting" parts or clusters is called *clustering*. There exist several successful methods that come from the area of statistics [3, 4, 7]. These methods require that the feature space in which the clustering is performed is a metric space. The "interestingness" of the clustering for these methods is related to some measure of distance defined on the feature space.

Problems arise when some or all dimensions of the feature space do not have a metric defined on them. For such spaces it is often difficult to perform unsupervised clustering, and if it is done, the clustering criteria are often problem specific. However, additional help is often available in the form of suitable metrics defined on some of the dimensions of the feature space. It would be a goal to devise general hybrid methods that could take advantage of this.

2 Preliminaries

Rough set theory [5] deals with the notion of set approximations. The members of sets are objects which are accessible only through the observations recorded about them. These observations are called *attributes* and their values are recorded in an *information system* [6].

Given the set of objects of the universe $\mathcal{U}$, an attribute a is a mapping $a : \mathcal{U} \to V_a$ where V_a is the attribute value set. An information system is a tuple $\mathbb{A} = (\mathcal{U}, \mathcal{A})$ of a universe of objects and a set of attributes $\mathcal{A}$. Let $x \in \mathcal{U}$ and let $\mathcal{B}$ be a set of attributes on $\mathcal{U}$. Then we define the *information vector* of x wrt. $\mathcal{B}$ as $Inf_{\mathcal{B}}(x) = \{(a, a(x)) \mid a \in \mathcal{B}\}$. Two objects $x, y \in \mathcal{U}$ are said to be *indiscernible* wrt. to $\mathcal{B}$ if and only if $Inf_{\mathcal{B}}(x) = Inf_{\mathcal{B}}(y)$.

A *reduct* $\mathbb{B}$ of an information system $\mathbb{A} = (\mathcal{U}, \mathcal{A})$ is an information system $\mathbb{B} = (\mathcal{U}, \mathcal{A}')$ where $\mathcal{A}'$ is a minimal subset of $\mathcal{A}$ such that all indiscernible objects of $\mathcal{U}$ wrt. $\mathcal{A}$ still are indiscernible wrt. $\mathcal{A}'$. The set of reducts of $\mathbb{A}$ is denoted $RED(\mathbb{A})$. Computing reducts from information systems is equivalent to finding prime implicants of a Boolean equation [1]. This problem is considered to be NP-hard. Using Boolean reasoning to find reducts of information systems is discussed amongst others in [8].

The *indiscernibility relation* $\sim_{\mathcal{B}}$ on $\mathcal{U}$ given by $\mathcal{B}$, is defined as $\forall x, y \in \mathcal{U}$, $x \sim_{\mathcal{B}} y \Leftrightarrow Inf_{\mathcal{B}}(x) = Inf_{\mathcal{B}}(y)$. This relation is an equivalence relation. The equivalence class to which an element $x \in \mathcal{U}$ belongs to is denoted as $[x]_{\mathcal{B}}$. Letting $X \subseteq \mathcal{U}$, we define the *upper approximation* of X wrt. $\mathcal{B}$ as $\overline{X}_{\mathcal{B}} = \{x \in \mathcal{U} \mid [x]_{\mathcal{B}} \cap X \neq \emptyset\}$ and the *lower approximation* of X wrt. $\mathcal{B}$ as $\underline{X}_{\mathcal{B}} = \{x \in \mathcal{U} \mid [x]_{\mathcal{B}} \cap X = [x]_{\mathcal{B}}\}$. The rough *membership function* $\mu_X^{\mathcal{B}} : \mathcal{U} \to \mathbb{R}$ of an element $x \in \mathcal{U}$ in a set $X \subseteq \mathcal{U}$ is defined as $\mu_X^{\mathcal{B}}(x) = \frac{|[x]_{\mathcal{B}} \cap X|}{|[x]_{\mathcal{B}}|}$. Let $\mathbb{X} = \{X_i \in \mathbf{P}(\mathcal{U}) | i \in \mathbf{I}\}$ for some non-empty index set $\mathbf{I}$ where $|\mathbf{I}| = n$ and $\mathbf{P}(\mathcal{U})$ is the power-set of $\mathcal{U}$. Let $X^1 = X$ and $X^0 = \mathcal{U} - X$ and let

$\mathbb{X}^v = \bigcap_j X_j^{i_j}$ where $v = (i_1 i_2 \cdots i_n) \in \{0,1\}^n$. Now define $AT(\mathbb{X}) = \{\mathbb{X}^v \mid v \in \{0,1\}^n \wedge \mathbb{X}^v \neq \emptyset\}$. The elements of $AT(\mathbb{X})$ are called *atoms* of $\mathbb{X}$.

Proposition 1 *Let $\mathbb{X} = \{X_i \in \mathbf{P}(\mathcal{U}) \mid i \in \mathbf{I}\}$ for some index set $\mathbf{I}$. The atoms of $\mathbb{X}$ constitute a partition of $\mathcal{U}$.*

Proposition 2 *Let $\mathbb{X}$ be a set of pairwise disjoint elements of $\mathbf{P}(\mathcal{U})$. The rough membership function of the union of elements of $\mathbb{X}$ wrt. the attribute set $\mathcal{B}$ on $\mathcal{U}$ defined as*

$$\mu^{\mathcal{B}}_{\cup\mathbb{X}}(x) = \sum_{X \in \mathbb{X}} \mu^{\mathcal{B}}_X(x) \text{ for any } x \in \mathcal{U}$$

is well-defined.

Proposition 2 ensures that one can calculate the rough membership degree of an object in any union of any sets by summation over the atoms generated by the sets in the union.

3 Methodology

We will regard a *clustering* of a universe $\mathcal{U}$ as a set

$$\mathbf{C} = \{C_i \in \mathbf{P}(\mathcal{U}) \mid \bigcup_i C_i = \mathcal{U}\}$$

where $C_k \subseteq C_j \Leftrightarrow k = j$. The elements of the clustering should reflect some "similarity" criterion on the objects of the universe. We might express this as two criteria that a clustering should meet: The inter-cluster "distance" should be maximal, and intra-cluster "variance" should be minimal. As some or all of the dimensions of the feature space lack a metric, we must discover a meaningful measure of similarity, and once this is found we will search for a clustering as a partition $\mathbb{P}_{\mathcal{B}'}$ of the universe induced by the equivalence relation $\sim_{\mathcal{B}'}$.

3.1 Valuation of attributes

In this section we propose a measure of the importance of an attribute for the classification of objects into elements of a clustering.

Definition 1 *Let $\mathbf{C}$ be a clustering of $\mathcal{U}$ and let $\mathcal{B}$ be a non-empty attribute set on $\mathcal{U}$. We then define*

- *The* upper crispness *of $\mathbf{C}$ wrt. to the attribute set $\mathcal{B}$ as*

$$\mathbf{Cr}^+_{\mathcal{B}}(\mathbf{C}) = \frac{1}{|\mathcal{U}|} \sum_{x \in \mathcal{U}} \max_{Y \in \mathbf{C}} (\mu^{\mathcal{B}}_Y(x))$$

- *The* lower crispness *of $\mathbf{C}$ wrt. to the attribute set $\mathcal{B}$ as*

$$\mathbf{Cr}^-_{\mathcal{B}}(\mathbf{C}) = \frac{1}{|\mathcal{U}|} \sum_{x \in \mathcal{U}} \min_{Y \in \mathbf{C}} (\mu^{\mathcal{B}}_Y(x))$$

  where only those Y such that $\mu^{\mathcal{B}}_Y(x) > 0$ are considered in the summation.

- *The* crispness *of $\mathbf{C}$ wrt. to the attribute set $\mathcal{B}$ as the interval*

$$\mathbf{Cr}_{\mathcal{B}}(\mathbf{C}) = [\mathbf{Cr}^-_{\mathcal{B}}(\mathbf{C}), \mathbf{Cr}^+_{\mathcal{B}}(\mathbf{C})]$$

The size of the crispness interval can be used as a measure of certainty of overall classification of the objects into the members of the clustering $\mathbf{C}$. Using interval calculus we can define the value of an attribute.

Definition 2 *Let $\mathcal{B}$ be an attribute set on the universe $\mathcal{U}$, and let $\mathbf{C}$ be a clustering of $\mathcal{U}$. The value of an attribute a for the crispness of $\mathbf{C}$ is defined as*

$$val^{\mathbf{C}}_{\mathcal{B}}(a) = \begin{cases} \mathbf{Cr}_{\mathcal{B}}(\mathbf{C}) - \mathbf{Cr}_{\mathcal{B}-\{a\}}(\mathbf{C}) & \text{if } |\mathcal{B}| > 1 \\ \mathbf{Cr}_{\mathcal{B}}(\mathbf{C}) & \text{if } \mathcal{B} = \{a\} \\ 0 & \text{otherwise} \end{cases}$$

Proposition 3 *Let $\mathbb{A} = (\mathcal{U}, \mathcal{A})$ be an information system. Let $\mathcal{B} \subseteq \mathcal{A}$, and let $\mathbb{P}_{\mathcal{B}}$ be the partition of the universe given by the indiscernibility relation $\sim_{\mathcal{B}}$ on $\mathcal{U}$. Then*

$$val^{\mathbb{P}_{\mathcal{B}}}_{\mathcal{B}}(a) \neq [0,0] \Leftrightarrow a \in \bigcap_{(\mathcal{U}, A') \in RED((\mathcal{U},\mathcal{B}))} A'$$

This proposition forces us to work with reducts only, as we otherwise cannot guarantee that there are any attributes which have a value other than $[0,0]$.

3.2 Distance function

In this section we propose a semi-metric on the feature space using the value of attributes proposed above.

Definition 3 *Let a be an attribute on $\mathcal{U}$ and let $x, y \in \mathcal{U}$. If there exists a metric m_{V_a} on V_a we define*

$$\mathbf{N}_{V_a}(a(x), a(y)) = \frac{1}{M} m_{V_a}(a(x), a(y))$$

where $M = \max(m_{V_a}(a(w), a(z))) \mid w, z \in \mathcal{U}$. If there does not exist a metric on V_a we define instead

$$\mathbf{N}_{V_a}(a(x), a(y)) = F(|V_a|)$$

where F is some mapping $F : \mathbb{N} \to \mathbb{R}$.

Definition 4 *Let $\mathcal{B}$ be an attribute set on the universe $\mathcal{U}$, let $\mathbb{Y}$ be a set of reducts of $(\mathcal{U}, \mathcal{B})$ and let $\mathbb{P}_{\mathcal{B}}$ be the partition of $\mathcal{U}$ given by the indiscernibility relation $\sim_{\mathcal{B}}$ on $\mathcal{U}$. Then we define the* clustering distance $\mathbf{Cd}_{\mathcal{B}}^{\mathbb{Y}} : \mathbb{P}_{\mathcal{B}} \times \mathbb{P}_{\mathcal{B}} \to \mathbb{R}$ *as: for all $x, y \in \mathbb{P}_{\mathcal{B}}$*

$$\mathbf{Cd}_{\mathcal{B}}^{\mathbb{Y}}(x, y) = \sum_{(\mathcal{U},\mathcal{Y})\in\mathbb{Y}} \sum_{a\in D_{\mathcal{Y}}(x,y)} \mathbf{N}_{V_a}(a(x), a(y)) \lceil val_{\mathcal{Y}}^{\mathbb{P}_{\mathcal{Y}}}(a) \rceil$$

where $D_{\mathcal{Y}}(x, y) = \{a \in \mathcal{Y} \mid a(x) \neq a(y)\}$ and $\lceil val_{\mathcal{Y}}^{\mathbb{P}_{\mathcal{Y}}}(a) \rceil$ is the upper limit of the value interval.

It is not difficult to convince oneself that the clustering distance does not fulfill the triangle inequality, thus the clustering distance is only a semi-metric [10], a result which follows from the following theorem.

Theorem 1 *Let $\mathcal{B}$ be an attribute set on $\mathcal{U}$ and let $\mathbb{Y}$ be a non-empty set of reducts of $(\mathcal{U}, \mathcal{B})$. If $F(x) > 0 \ \forall x \in \mathbb{Z}^+$, where F is the mapping in definition 3, then $\mathbf{Cd}_{\mathcal{B}}^{\mathbb{Y}}$ is a semi-metric on $\mathbb{P}_{\mathcal{B}}$, the partition of $\mathcal{U}$ induced by $\mathcal{B}$.*

3.3 Constructing metrics

In [9, 8] a metric based on a reflexive and symmetric *tolerance relation* τ is described. Informally, this metric is the same as the cost of the shortest path between two vertices in the graph with vertices $\mathcal{U}$ and edges given by τ and where the edges have unit cost. Using the idea of constructing undirected graphs where $\mathcal{U}$ is the set of vertices, we can construct several metrics, informally presented as:

- Construct a (complete) graph with vertices $\mathcal{U}$, and edges with cost of the clustering distance function described in section 3.2. Use the shortest path between two nodes as a metric.
- Construct a (complete) graph with vertices $\mathcal{U}$, and edges with cost of the clustering distance function described in section 3.2. Calculate a minimal spanning tree and use this to construct a metric.
- Construct a graph with $\mathcal{U}$ as vertices, and add an edge to the graph if the clustering distance function value is below some threshold. Use this graph to construct τ and use the tolerance metric m_τ. Here we must take into account that the resulting graph might not be connected.

3.4 Clustering strategy

With a metric defined on the entire feature space, we can use most of the clustering strategies/algorithms that stem from the field of statistics. We will here describe a simple strategy that results in the induced partition we are searching for.

A clustering is often evaluated according to whether the inter-cluster "distance" is "large" and the "distance" between two elements in the same cluster is "small". This can be formulated as the problem of minimizing a *criteria function* which is a formal representation of an evaluation mechanism of a clustering. A variation of the above is a threshold-based clustering scheme based on a distance measure and two thresholds, one for distance between clusters, and one for distance within the clusters [7, 8].

When a metric m has been chosen, a criteria function can be constructed. The criteria function should reflect the similarity criteria presented above. A candidate is:

$$J(\mathbf{C}) = \frac{\text{average intra-cluster scatter}}{\text{average inter-cluster distance}}.$$

If we let $\alpha(C_i, C_j) = \min m(x, y) \mid x \in C_i, y \in C_j$ be the minimal distance between sets C_i and C_j in $\mathbf{C}$, and let

$$\beta(C_i) = \frac{2}{n(n-1)} \sum_{k=1}^{|C_i|-1} \sum_{l=k+1}^{|C_i|} m(x_k, x_l)$$

be the intra-set scatter of set C_i, we can formulate our criteria function as:

$$J(\mathbf{C}) = \frac{\sum_i^n \beta(C_i)}{\frac{2}{(n-1)} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \alpha(C_k, C_l)}$$

where $n = |\mathbf{C}|$. A possible clustering strategy that incorporates the above common criteria function, starts with a given partition of the universe, and collapses equivalence classes to form larger clusters. The candidate equivalence classes for this operation are chosen using the metric found above. A pseudo-code description of this strategy can be found in figure 1.

3.5 A reference-based classifier

A possible use for the metric used in section 3.4 is in a reference-based classifier for previously unencountered or incomplete information vectors.

Input: $\mathcal{U}$, $\mathcal{B}_0$, a metric m and a threshold value τ

$i \leftarrow 0, w \leftarrow \infty$;
while $(J(\mathbb{P}_{\mathcal{B}_i}) - w > \tau)$ **do begin**
  $w \leftarrow J(\mathbb{P}_{\mathcal{B}_i})$;
  $(x, y) \leftarrow \arg\min m(x, y) \mid x \neq y \wedge x, y \in \mathbb{P}_{\mathcal{B}_i}$;
  Modify $\mathcal{B}_i$ such that (only) the equivalence classes of x and y are merged into one equivalence class, thus obtaining $\mathcal{B}_{i+1}$;
  $i \leftarrow i + 1$;
end

Figure 1: Pseudo-code of clustering strategy.

An example is a k-nearest neighbor (kNN) classifier [2, 7] which classifies the new, unlabeled information vector as having the class of a function of the k nearest neighbors. With missing data, one might take into consideration the corresponding data of the nearest neighbor in the restriction of the feature space to the dimensions present in the vector.

4 Summary

A method employing techniques from rough set theory has been presented for evaluating how important an attribute in an information system is wrt. assigning an object into an element of a clustering. This attribute evaluation mechanism (together with possible extra information about the value space of each attribute in the form of an existing metric on this space) is used to construct a semi-metric on the entire feature space. In order to arrive at a proper metric, possible combinations with techniques from graph theory are suggested. Furthermore, two potential applications have been proposed – clustering and construction of a reference-based classifier.

References

[1] F. M. Brown. *Boolean Reasoning.* Kluwer Academic Publishers, 1990.

[2] E. J. Caccio, S. M. Dunn, and M. Akay. Biosignal pattern recognition and interpretation systems. *IEEE Engineering in Medicine and Biology*, September 1993.

[3] I. Gath and A. B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-11(7), July 1989.

[4] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9), September 1990.

[5] Z. Pawlak. Rough sets. *International Journal of Information and Computer Science*, 11, 1982.

[6] Z. Pawlak. *Rough Sets, Theoretical Aspects of Reasoning about Data*, volume 9 of *Series D: System Theory, Knowledge Engineering and Problem Solving.* Kluwer Academic Publishers, 1991.

[7] R. Schalkoff. *Pattern Recognition, Statistical, Structural and Neural Approaches.* John Wiley & Sons, Inc., 1992.

[8] A. Skowron. Synthesis of adaptive decision systems from experimental data. In A. Aamodt and J. Komorowski, editors, *SCAI - 95 Fifth Scandinavian Conference on Artificial Intelligence*, volume 28 of *Frontiers in Artificial Intelligence and Applications.* IOS Press, 1995.

[9] A. Skowron, L. Polkowski, and J. Komorowski. Learning tolerance relations by boolean descriptors: Automatic feature extraction from data tables. In *The Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery*, November 1996.

[10] Eduard Čech. *Topological Spaces.* John Wiley & Sons, 1966.