Learning Predictive Clustering Rules

Bernard Ženko[1], Sašo Džeroski[1], and Jan Struyf[2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia
Bernard.Zenko@ijs.si, Saso.Dzeroski@ijs.si
[2] Department of Computer Science, Katholieke Universiteit Leuven, Belgium
Jan.Struyf@cs.kuleuven.be

Abstract. The two most commonly addressed data mining tasks are predictive modelling and clustering. Here we address the task of predictive clustering, which contains elements of both and generalizes them to some extent. Predictive clustering has been mainly evaluated in the context of trees. In this paper, we extend predictive clustering toward rules. Each cluster is described by a rule and different clusters are allowed to overlap since the sets of examples covered by different rules do not need to be disjoint. We propose a system for learning these predictive clustering rules, which is based on a heuristic sequential covering algorithm. The heuristic takes into account both the precision of the rules (compactness w.r.t. the target space) and the compactness w.r.t. the input space, and the two can be traded-off by means of a parameter. We evaluate our system in the context of several multi-objective classification problems.

1 Introduction

Predictive modeling or supervised learning aims at constructing models that can predict the value of a target attribute (dependent variable) from the known values for a set of input attributes (independent variables). A wide array of predictive modeling methods exist, which produce more or less (or not at all) interpretable models. Typical representatives of the group of methods that result in understandable and interpretable models are decision tree learning [14] and rule learning [7].

Clustering [9], on the other hand, is an unsupervised learning method. It tries to find subgroups of examples or clusters with homogeneous values for all attributes, not just the target attribute. In fact, the target attribute is usually not even defined in a clustering task. The result is a set of clusters and not necessarily their descriptions or models; usually we can link new examples to the constructed clusters based on e.g., proximity in the attribute space.

Predictive modeling and clustering are therefore regarded as quite different techniques. Nevertheless, different viewpoints also exist [10] which stress the many similarities that some predictive modeling techniques, most notably techniques that partition the example space, such as decision trees, share with clustering. Decision trees partition the set of examples into subsets with homogeneous values for *the target attribute*, while clustering methods search for subsets in which the examples have homogeneous values for *all the attributes*.

In this paper, we consider the task of predictive clustering [1, 2], which contains elements of both predictive modelling and clustering and generalizes them to some extent. In predictive clustering, one can simultaneously consider homogeneity along the target attribute and the input attributes, and trade-off one for the other. It has been argued [1] that predictive clustering is useful in noisy domains and in domains with missing values for the target attribute. Furthermore, predictive clustering has been proven useful in applications with non-trivial targets such as multi-objective classification and regression [2, 17], ranking [20], and hierarchical multi-classification [18].

Predictive clustering has been evaluated mainly in the context of trees. In this paper we extend predictive clustering toward rules. We call the resulting framework predictive clustering rules (PCRs). The task of learning PCRs generalizes the task of rule induction, on one hand, and clustering, and in particular item set constrained clustering [15, 16], on the other.

Since learning PCRs is a form of constrained clustering, it is directly related to constraint-based data mining and inductive databases (IDBs). Constraint-based clustering is an under-researched topic in constraint-based data mining and the present research is a step towards rectifying this. Bringing the two most common data mining tasks closer together (as done in predictive clustering) moves us towards finding a general framework for data mining, which is also the main goal of IDBs.

The rest of this paper is organized as follows. In the next section, we discuss prediction, clustering and predictive clustering in more detail. Section 3 extends predictive clustering toward rules and proposes the first system for building predictive clustering rules. The algorithm used in the system is a heuristic sequential covering algorithm and the heuristic trades-off homogeneity w.r.t. the target attributes and homogeneity w.r.t. the input attributes. We compare our system to related approaches in Section 4. Section 5 evaluates the system on a number of multi-objective classification and regression data sets. The paper ends with a discussion of further work and a conclusion.

2 Prediction, Clustering, and Predictive Clustering

The tasks of predictive modelling and clustering are two of the oldest and most commonly addressed tasks in data analysis and data mining. Here we briefly introduce each of them and discuss predictive clustering, a task that combines elements of both prediction and clustering.

2.1 Predictive Modelling

Predictive modeling aims at constructing models that can predict a target property of an object from a description of the object. Predictive models are learned from sets of examples, where each example has the form (D, T), with D being an object description and T a target property value. While a variety of languages ranging from propositional to first order logic have been used for D, T is almost always considered to consist of a single target attribute called the class: if this attribute is discrete we are dealing with a classification problem and if continuous with a regression problem.

In practice, D is most commonly a vector and each element of this vector is the value for a particular attribute (attribute-value representation). In the remainder of the paper, we will consider both D and T to be vectors of attribute values (discrete or real-valued). If T is a vector with several target attributes, then we call the prediction task multi-objective prediction. If T only contains discrete attributes we speak of multi-objective classification. If T only contains continuous attributes we speak of multi-objective regression.

Predictive models can take many different forms that range from linear equations to logic programs. Two commonly used types of models are decision trees [14] and rules [7]. Unlike (regression) equations that provide a single predictive model for the entire example space, trees and rules divide the space of examples into subspaces and provide a simple prediction or predictive model for each of these.

2.2 Clustering and Clustering Trees

Clustering [9] in general is concerned with grouping objects into classes of similar objects. Given a set of examples (object descriptions), the task of clustering is to partition these examples into subsets, called clusters. Note that examples do not contain a target property to be predicted, but only an object description (which is typically a vector of attribute-values D). The goal of clustering is to achieve high similarity between objects within individual clusters and low similarity between objects that belong to different clusters.

Conventional clustering focuses on distance-based cluster analysis. The notion of a distance (or conversely, similarity) is crucial here: examples are considered to be points in a metric space (a space with a distance measure). A prototype (or prototypical example) may be used as a representative for a cluster. Usually, the prototype is the point with the lowest average distance to all the examples in the cluster, i.e., the mean or the medoid of the examples.

In conceptual clustering [12], a symbolic representation of the resulting clusters is produced in addition to the partition into clusters: we can thus consider each cluster to be a concept (much like a class in classification). In this context, a decision tree structure may be used to represent a hierarchical clustering: such a tree is called a clustering tree [1]. In a clustering tree each node represents a cluster. The conjunction of conditions on the path from the root to that node gives a symbolic representation of the cluster. Essentially, each cluster has a symbolic description in the form of a rule (IF conjunction of conditions THEN cluster), while the tree structure represents the hierarchy of clusters. Clusters that are not on the same branch of a tree do not overlap.

Given the above, predictive modelling approaches which divide the set of examples into subsets, such as decision tree and rule induction, are in a sense very similar to clustering. A major difference is that they partition the space

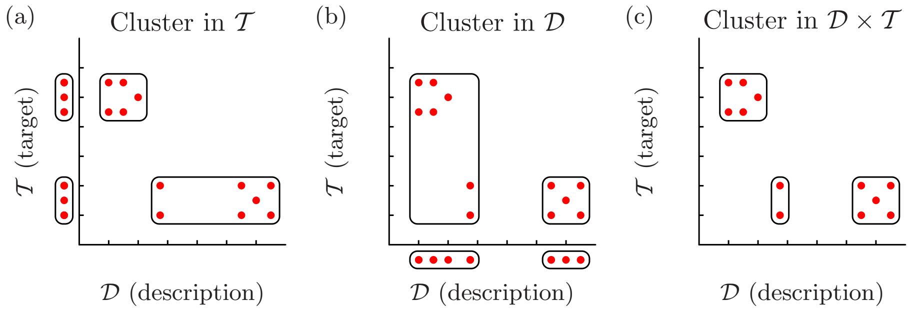

(a)
Cluster in $\mathcal{T}$
$\mathcal{T}$ (target)
$\mathcal{D}$ (description)
(b)
Cluster in $\mathcal{D}$
$\mathcal{T}$ (target)
$\mathcal{D}$ (description)
(c)
Cluster in $\mathcal{D} \times \mathcal{T}$
$\mathcal{T}$ (target)
$\mathcal{D}$ (description)

Fig. 1. Predictive modelling (a), clustering (b), and predictive clustering (c).

of examples into subsets with homogeneous values of the *target attribute*, while (distance-based) clustering methods seek subsets with homogeneous values of the *descriptive attributes*. This is illustrated in Fig. 1. Assume that each example has a description $D \in \mathcal{D}$ and is labeled with a target value $T \in \mathcal{T}$. A predictive tree learner will build a tree with leaves that are as pure as possible w.r.t. the target value, i.e., it will form clusters that are homogeneous in $\mathcal{T}$, as shown in Fig. 1.a. The reason is that the quality criterion that is used to build the tree (e.g., information gain [14]) is based on the target attributes only[3]. In unsupervised clustering, on the other hand, there is no target value defined and the clusters that are generated will be homogeneous w.r.t. $\mathcal{D}$, as shown in Fig. 1.b. In the next section, we will consider predictive clustering, which in general searches for clusters that are homogeneous w.r.t. to both $\mathcal{D}$ and $\mathcal{T}$ (Fig. 1.c).

2.3 Predictive Clustering

The task of predictive clustering [2] combines elements from both prediction and clustering. As is common in clustering, we seek clusters of examples that are similar to each other (and dissimilar to examples in other clusters), but in general taking both the descriptive and the target attributes into account. In addition, a predictive model must be associated with each cluster; the model gives a prediction of the target variables T in terms of the attributes D for all examples that are established to belong to that cluster.

In the simplest and most common case, the predictive model associated to a cluster would be the projection on T of the prototype of the examples that belong to that cluster. This would be a simple average when T is a single continuous variable. In the discrete case, it would be a probability distribution across the discrete values or the mode thereof. When T is a vector, the prototype would be a vector of averages and distributions/modes.

[3] Because the leaves of a decision tree have conjunctive descriptions in $\mathcal{D}$, the corresponding clusters will also have some homogeneity w.r.t. $\mathcal{D}$, but the latter is not optimized by the system.

To summarize, in predictive clustering, each cluster has both a symbolic description (in terms of a language bias over D) and a predictive model (a prototype in T) associated to it, i.e., the resulting clustering is defined by a symbolic model. If we consider a tree based representation, then this model is called a predictive clustering tree. Predictive clustering trees have been proposed by Blockeel et al. [2]. In the next section, we will propose predictive clustering rules, a framework in which the clustering model is represented as a rule set.

3 Predictive Clustering Rules (PCRs)

This section presents the main contribution of this paper, which is the predictive clustering rules (PCRs) framework. We start with a general definition of PCRs. Then we apply this general definition to the multi-objective prediction setting. Finally, we propose a system for learning PCRs in this setting.

3.1 Definition

The task of learning a set of PCRs is defined as follows.

Given:

- a target space $\mathcal{T}$
- a description space $\mathcal{D}$
- a set of examples $E = \{e_i\}$, with $e_i \in \mathcal{D} \times \mathcal{T}$
- a declarative language bias B over $\mathcal{D}$
- a distance measure d that computes the distance between two examples
- a prototype function p that computes the prototype of a set of examples

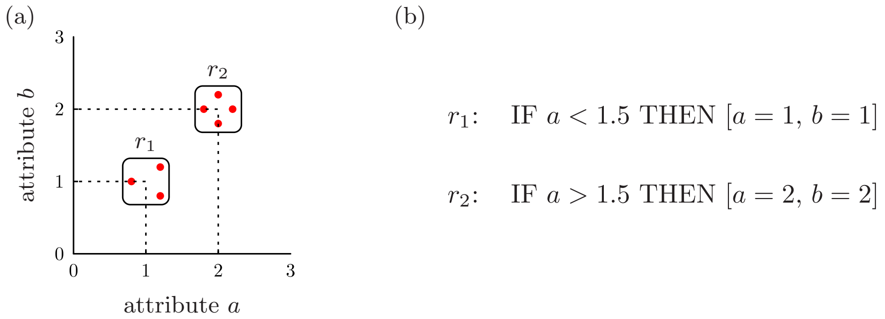
Find a set of clusters, where

- each cluster is associated with a description expressed in B
- each cluster has an associated prediction expressed as a prototype
- within-cluster distance is low (similarity is high) and
- between-cluster distance is high (similarity is low)

Each cluster can thus be represented as a so-called predictive clustering rule, which is a rule of the form “IF cluster description THEN cluster prototype”.

Example 1. Consider the data set shown in Fig 2.a. It has two numeric attributes: a is a descriptive attribute and b is the target attribute, i.e., $\mathcal{D} = \mathcal{T} = \mathbb{R}$. Suppose that the distance metric is the Euclidean distance over $\mathbb{R}^2$. The corresponding prototype is the vector average. If the language bias B allows conjunctions of tests comparing a to a particular constant, then a possible set of PCRs for this data set is shown in Fig 2.b.

Note that the description in the conditional part of a PCR only takes D into account and not T. The reason is that it must be possible to apply the rule to unseen examples later on for which T is not defined.



r_1: IF $a < 1.5$ THEN $[a = 1,\ b = 1]$

r_2: IF $a > 1.5$ THEN $[a = 2,\ b = 2]$

Fig. 2. A data set (left) and the corresponding set of predictive clustering rules (right).

There are two main differences between PCRs and predictive clustering trees. The first difference is that predictive clustering trees represent a hierarchical clustering of the data whereas the clustering corresponding to a set of PCRs is flat. The other difference is that the clusters defined by a set of PCRs may overlap. In fact, there are two possible interpretations of a set of PCRs: the rules can be ordered and treated as a decision list. In that case, a given example belongs to the cluster of the first rule in the list that fires and the resulting clustering is disjoint. On the other hand, if the rules are considered unordered, then the clusters may overlap as several rules may apply to a given example. If in the latter case a set of rules fire for a given example, then a method is required for combining the predictions of the different rules. Such a method is not required in predictive clustering trees since in that case the clusters are guaranteed to be disjoint. We will propose a suitable combining method later in the paper.

3.2 PCRs for Multi-Objective Prediction (MOP)

In this section, we discuss the PCR framework in the context of multi-objective prediction (MOP) tasks. This includes multi-objective classification and multi-objective regression as discussed before. As a result, the examples will be of the form (D, T) with D and T both vectors of attribute-values. MOP has two main advantages over using a separate model for each target attribute: (1) a single MOP model is usually much smaller than the total size of the individual models for all attributes, and (2) a MOP model may explicitate dependencies between the different target variables [17].

The distance metric that we will use in the clustering process takes both D and T into account and is defined as follows.

$$d = (1 - \tau) d_D + \tau d_T. \tag{1}$$

It has two components, one for the descriptive part d_D and a second one for the target part d_T and the relative contribution of the two components can be changed by means of the parameter τ.

In our rule induction algorithm, we will estimate the quality of a (partial) rule that covers a set of examples S as the average distance of an example in S to the prototype of S. We will call this the "compactness" of the rule. Because the attributes can in general be nominal or numeric, different measures for each type are needed which are then combined (added) into a single measure.

For nominal attributes, the prototype is a vector with as components the frequencies of each of the attribute's values in the given set of examples S, i.e., for an attribute with k possible values (v_1 to v_k), the prototype is of the form $[f_1, f_2, \ldots, f_k]$, with f_i the frequency of v_i in S. The distance between an example and the prototype is defined as follows: if the attribute value for the example is v_i, then the distance to the prototype is defined as $(1 - f_i)$. For numeric attributes, the prototype is the mean of the attribute's values and the distance between an example and the prototype is computed as the absolute difference. Numeric attributes are normalized during a preprocessing step such that their mean is zero and their variance is one.

Example 2. Consider a data set with a nominal attribute a with possible values $\oplus$ and $\ominus$ and a numeric attribute b. There are three examples in S: $[\oplus, 1]$, $[\oplus, 2]$, and $[\ominus, 1]$. The prototype for a is the vector $[2/3, 1/3]$ and the prototype for b is 2. The compactness of a is $4/9$ and the compactness of b is $2/3$. The combined compactness is $2/3 + 4/9 = 10/9$.

Note that our compactness measure is actually an "incompactness" measure, since smaller values mean more compact sets of examples.

The declarative bias will restrict our hypothesis language to rules consisting of conjunctions of attribute-value conditions over the attributes D. In particular, we consider subset tests for nominal attributes and inequality tests for numeric attributes. Additional language constraints are planned for consideration.

3.3 Learning Predictive Clustering Rules

This section describes our system for learning PCRs. The majority of rule induction methods are based on a sequential covering algorithm and among these the CN2 algorithm [4] is well known. Our system is based on this algorithm, but several important parts are modified. In this section we first briefly describe the original CN2 algorithm, and then we present our modifications.

Rule Induction with CN2 The CN2 algorithm iteratively constructs rules that cover examples with homogeneous target variable values. The heuristic used to guide the search is simply the accuracy of the rule under construction. After a rule has been constructed, the examples covered by this rule are removed from the training set, and the procedure is repeated on the new data set until the data set is empty or no new rules are found. The rules constructed in this way are ordered, meaning that they can be used for prediction as a decision list; we test rules on a new example one by one and the first rule that fires is used for prediction of the target value of this example. Alternatively, CN2 can also construct unordered rules if only correctly classified examples are removed from the training set after finding each rule and if rules are built for each class in turn. When using unordered rules for prediction, several rules can fire on each example and a combining method is required as discussed before.

The Search Heuristic: Compactness The main difference between CN2 and the approach presented in this paper is the heuristic that is used for guiding the search for rules. The purpose of the heuristic is to evaluate different rules; it should measure the quality of each rule separately and/or the quality of the whole rule set.

One of the most important properties of rules (and other models) is their accuracy, and standard CN2 simply uses this as a heuristic. Accuracy is only connected to the target attribute. Our goal when developing predictive clustering rules was (besides accuracy) that the induced rules should cover compact subsets of examples, just as clustering does. For this purpose we need a heuristic which takes into account the target attributes as well as the descriptive attributes.

As explained above, we will use the compactness (average distance of an example covered by a rule to the prototype of this set of examples). The compactness takes into account both the descriptive and the target attributes and is a weighted sum of the compactness along each of the dimensions (the latter are normalized to be between 0 and 1). At present only a general weight τ is applied for putting the emphasis on the targets attributes ($\tau = 1$) or the input attributes ($\tau = 0$): target attributes should in general have higher weights in order to guide the search toward accurate rules.

Weighted Covering The standard covering algorithm removes the examples covered by a rule from the training set in each iteration. As a consequence, subsequent rules are constructed on smaller example subsets which can be improperly biased and can have small coverage. To overcome these shortages we employ the weighted covering algorithm [11]. The difference is that once an example is covered by a new rule, it is not removed from the training set but instead, its weight is decreased. As a result, the already covered example will be less likely covered in the next iterations. We use the additive weighting scheme, which means that the weight of an example after being covered m times is equal to $\frac{1}{1+m}$. Finally, when the example is covered more than a predefined number of times (in our experiments five), the example is completely removed from the training set.

Probabilistic Classification As already mentioned, the original CN2 algorithm can induce ordered or unordered rules. In case of ordered rules (i.e., a decision list) the classification is straightforward. We scan the rules one by one and whichever rule fires first on a given example is used for prediction. If no rule fires, the default rule is used. When classifying with unordered rules, CN2 collects class distributions of all rules that fire on an example and uses them for weighted voting. We use the same probabilistic classification scheme even though our unordered rules are not induced for each possible class value separately.

4 Related Work

Predictive modeling and clustering are regarded as quite different tasks. While there are many approaches addressing each of predictive modelling and clustering, few approaches look at both or try to relate them. A different viewpoint is taken by Langley [10]: predictive modeling and clustering have many similarities and this has motivated some recent research on combining prediction and clustering.

The approach presented in this paper is closely related to predictive clustering trees [2], which also address the task of predictive clustering. The systems TILDE [2] and CLUS [3] use a modified top-down induction of decision trees algorithm to construct clustering trees (which can predict values of more than one target variables simultaneously). So far, however, distances used in TILDE and CLUS systems have considered attributes or classes separately, but not both together, even though the idea was presented in [1].

Our approach uses a rule-based representation for predictive clustering. As such, it is closely related to approaches for rule induction, and among these in particular CN2 [4]. However, it extends rule induction to the more general task of multi-objective prediction. While some work exists on multi-objective classification with decision trees (e.g., [19]), the authors are not aware of any work on rule-induction for multi-objective classification. Also, little work exists on rule-based regression (e.g., $R2$ [21] for propositional learning and FORS [8] for first order logic learning), let alone rule-based multi-objective regression (or multi-objective prediction in general, with mixed continuous and discrete targets).

Related to rule induction is subgroup discovery [11], which tries to find and describe interesting groups of examples. While subgroup discovery algorithms are similar to rule induction ones, they have introduced interesting innovations, including the weighted covering approach used in our system.

Another related approach to combining clustering and classification is *itemset constrained clustering* [15, 16]. Here the attributes describing each example are separated in two groups, called feature items and objective attributes. Clustering is done on the objective attributes, but only clusters which can be described in terms of frequent item sets (using the feature items attributes) are constructed. As a result each cluster can be classified by a corresponding frequent item set.

As in our approach, itemset classified clustering tries to find groups of examples with small variance of the objective attributes. As compared to itemset classified clustering, our approach allows both discrete (and not only binary attributes / items) and continuous variables on the feature/attribute side, as well as the objective/target side. Itemset constrained clustering is also related to subgroup discovery, as it tries to find interesting groups of examples, rather than a set of (overlapping) clusters that cover all examples. A second important difference is that in itemset classified clustering the distance metric takes only the objective attributes into account, whereas the rules constructed by our approach are also compact w.r.t. the descriptive space.

Table 1. The attributes of the river water quality data set.

Independent attributes	Target attributes
physical & chemical properties	taxa – presences/absences
numeric type	nominal type (0,1)
water temperature	Cladophora sp.
alkalinity (pH)	Gongrosira incrustans
electrical conductivity	Oedogonium sp.
dissolved O_2	Stigeoclonium tenue
O_2 saturation	Melosira varians
CO_2 conc.	Nitzschia palea
total hardness	Audouinella(Chantransia) chalybea
NO_2 conc.	Erpobdella octoculata
NO_3 conc.	Gammarus fossarum
NH_4 conc.	Baetis rhodani
PO_4 conc.	Hydropsyche sp.
Cl conc.	Rhyacophila sp.
SiO_2 conc.	Simulium sp.
chemical oxygen demand – $KMnO_4$	Tubifex sp.
chemical oxygen demand – $K_2Cr_2O_7$	
biological oxygen demand (BOD)	

5 Experiments

The current implementation of predictive clustering rules has been tested on several classification problems with multiple target attributes. For each data set two sets of experiments have been performed. First, we tried to test the performance of our method when predicting multiple target attributes at once in comparison to single target attribute prediction task. In the second set of experiments we investigated the influence of the target weighting parameter (τ) on the accuracy and compactness of induced rules.

5.1 Data Sets

There are not a lot of publicly available data sets suitable for multi-target classification. However, some of the data sets from the UCI repository [13] can also be used for this purpose, namely the data sets *monks, solar-flare,* and *thyroid*. The first two data sets have three target attributes each, while the third has seven.

In addition to these UCI data sets we have also used Slovenian rivers water quality data set (*water-quality*). The data set comprises biological and chemical data that were collected through regular monitoring of rivers in Slovenia. The data come from the Environmental Agency of the Republic of Slovenia that performs water quality monitoring for most Slovenian rivers and maintains a database of water quality samples. The data cover a six year period, from 1990 to 1995 and have been previously used in [5].

Table 2. Monks data set. Accuracies of predictive clustering rules (PCR) and predictive clustering trees (PCT) used for multi-objective prediction of all target attributes together and for single target prediction of each target attribute separately.

	PCR		PCT	
Target attribute	All	Indiv.	All	Indiv.
monk–1	0.803	0.810	0.711	0.764
monk–2	0.671	0.669	0.664	0.627
monk–3	0.935	0.935	0.972	0.972
Average accuracy	0.803	0.805	0.782	0.788

Table 3. Solar-flare data set. Accuracies of predictive clustering rules (PCR) and predictive clustering trees (PCT) used for multi-objective prediction of all target attributes together and for single target prediction of each target attribute separately.

	PCR		PCT	
Target attribute	All	Indiv.	All	Indiv.
class–c	0.828	0.829	0.829	0.826
class–m	0.966	0.966	0.966	0.966
class–x	0.995	0.995	0.995	0.995
Average accuracy	0.930	0.930	0.930	0.929

Biological samples are taken twice a year, once in summer and once in winter, while physical and chemical analysis are performed several times a year for each sampling site. The physical and chemical samples include the measured values of 15 different parameters. The biological samples include a list of all taxa (plant and animal species) present at the sampling site. All the attributes of the data set are listed in Table 1. In total, 1060 water samples are available in the data set. In our experiments we have considered the physical and chemical properties as independent attributes, and presences/absences of taxa as target attributes.

5.2 Results

The first set of experiments was performed in order to test the appropriateness of predictive clustering rules for multiple target prediction. In all experiments the minimal number of examples covered by a rule was 20, and the weight of target attributes (τ) was set to 1. The results of 10-fold cross validation can be seen for each data set separately in Tables 2, 3, 4, and 5. The first columns in tables are the accuracies for each target attribute as predicted by the PCR multi-target models and in the second columns are accuracies as predicted by the PCR single-target models. Third and fourth columns are accuracies for predictive clustering trees (PCT). The last rows in the tables give the average accuracies across all target attributes.

Table 4. Thyroid data set. Accuracies of predictive clustering rules (PCR) and predictive clustering trees (PCT) used for multi-objective prediction of all target attributes together and for single target prediction of each target attribute separately.

	PCR		PCT	
Target attribute	All	Indiv.	All	Indiv.
hyperthyroid	0.974	0.975	0.983	0.984
hypothyroid	0.941	0.947	0.989	0.989
binding protein	0.955	0.961	0.974	0.975
general health	0.970	0.972	0.984	0.985
replacement theory	0.961	0.963	0.985	0.990
antithyroid treatment	0.996	0.996	0.996	0.996
discordant results	0.979	0.979	0.987	0.989
Average accuracy	0.968	0.971	0.986	0.987

Looking at these average accuracies we can see that the performance of models predicting all classes together is comparable to the performance of single target models. There are no significant differences for the monks, solar-flare and thyroid data sets, while the multi-target model is somewhat worse than single-target models on the water quality data set. When comparing predictive clustering rules to predictive clustering trees, the performance of the latter is somewhat better on the thyroid and water quality data set but a little worse on the monks data set; there are no differences on the solar-flare data set.

Table 5. Water quality data set. Accuracies of predictive clustering rules (PCR) and predictive clustering trees (PCT) used for multi-objective prediction of all target attributes together and for single target prediction of each target attribute separately.

	PCR		PCT	
Target attribute	All	Indiv.	All	Indiv.
Cladophora sp.	0.594	0.629	0.630	0.648
Gongrosira incrustans	0.733	0.729	0.722	0.665
Oedogonium sp.	0.713	0.717	0.723	0.710
Stigeoclonium tenue	0.795	0.790	0.796	0.771
Melosira varians	0.569	0.611	0.638	0.643
Nitzschia palea	0.688	0.662	0.714	0.708
Audouinella chalybea	0.751	0.756	0.747	0.712
Erpobdella octoculata	0.721	0.741	0.712	0.691
Gammarus fossarum	0.628	0.654	0.664	0.688
Baetis rhodani	0.676	0.723	0.686	0.700
Hydropsyche sp.	0.584	0.604	0.614	0.630
Rhyacophila sp.	0.686	0.710	0.708	0.709
Simulium sp.	0.633	0.635	0.593	0.642
Tubifex sp.	0.728	0.745	0.735	0.739
Average accuracy	0.679	0.693	0.692	0.690

Table 6. Monks data set. The accuracy and cluster compactness of predictive clustering rules used for multiple target prediction of all target attributes together with different target attributes weightings (τ).

	τ					
Target attribute	0.5	0.7	0.8	0.9	0.95	1
monk–1	0.843	0.840	0.806	0.833	0.831	0.803
monk–2	0.671	0.671	0.671	0.671	0.671	0.671
monk–3	0.949	0.965	0.975	0.958	0.938	0.935
Average accuracy	0.821	0.826	0.817	0.821	0.813	0.803
Average compactness	0.487	0.486	0.486	0.495	0.506	0.516

Table 7. Solar flare data set. The accuracy and cluster compactness of predictive clustering rules used for multiple target prediction of all target attributes together with different target attributes weightings (τ).

	τ					
Target attribute	0.5	0.7	0.8	0.9	0.95	1
class–c	0.829	0.829	0.829	0.829	0.829	0.828
class–m	0.966	0.966	0.966	0.966	0.966	0.966
class–x	0.995	0.995	0.995	0.995	0.995	0.995
Average accuracy	0.930	0.930	0.930	0.930	0.930	0.930
Average compactness	0.158	0.159	0.161	0.181	0.207	0.239

The task of the second set of experiments was to evaluate the influence of the target weighting parameter (τ) on the accuracy and cluster compactness of induced rules (Tables 6, 7, 8, and 9). Rules were induced for predicting all target attributes together with six different values of the τ parameter. At the bottom of each table are the average accuracies of 10-fold cross-validation and average compactness of subsets of examples (clusters) covered by rules in each model.

Table 8. Thyroid data set. The accuracy and cluster compactness of predictive clustering rules used for multiple target prediction of all target attributes together with different target attributes weightings (τ).

	τ					
Target attribute	0.5	0.7	0.8	0.9	0.95	1
hyperthyroid	0.974	0.974	0.974	0.974	0.974	0.974
hypothyroid	0.927	0.927	0.927	0.927	0.928	0.941
binding protein	0.955	0.955	0.955	0.955	0.955	0.955
general health	0.938	0.938	0.938	0.938	0.939	0.970
replacement theory	0.961	0.961	0.961	0.961	0.961	0.961
antithyroid treatment	0.996	0.996	0.996	0.996	0.996	0.996
discordant results	0.979	0.979	0.979	0.979	0.979	0.979
Average accuracy	0.961	0.961	0.961	0.961	0.962	0.968
Average compompactness	1739	1797	1705	1591	1603	1605

Table 9. Water quality data set. The accuracy and cluster compactness of predictive clustering rules used for multiple target prediction of all target attributes together with different target attributes weightings (τ).

	τ					
Target attribute	0.5	0.7	0.8	0.9	0.95	1
Cladophora sp.	0.586	0.593	0.597	0.599	0.600	0.594
Gongrosira incrustans	0.733	0.733	0.733	0.733	0.733	0.733
Oedogonium sp.	0.716	0.719	0.716	0.717	0.718	0.713
Stigeoclonium tenue	0.792	0.793	0.793	0.793	0.793	0.795
Melosira varians	0.580	0.584	0.578	0.581	0.575	0.569
Nitzschia palea	0.656	0.664	0.657	0.655	0.661	0.688
Audouinella chalybea	0.753	0.753	0.753	0.753	0.753	0.751
Erpobdella octoculata	0.738	0.742	0.742	0.741	0.742	0.721
Gammarus fossarum	0.628	0.641	0.629	0.630	0.632	0.628
Baetis rhodani	0.676	0.676	0.676	0.676	0.676	0.676
Hydropsyche sp.	0.566	0.564	0.568	0.567	0.565	0.584
Rhyacophila sp.	0.684	0.685	0.685	0.685	0.685	0.686
Simulium sp.	0.639	0.640	0.640	0.640	0.646	0.633
Tubifex sp.	0.732	0.729	0.725	0.730	0.731	0.728
Average accuracy	0.677	0.680	0.678	0.679	0.679	0.679
Average compactness	0.348	0.348	0.348	0.348	0.348	0.350

The rules induced with larger weighting of the non-target attributes (smaller τ) are on average more compact on the monks and solar-flare data sets (smaller number for compactness means more compact subsets) while there is no clear trend for the thyroid data set and no influence at all on the water quality data set. Larger weighting of the non-target attributes has very little effect on the accuracy of the models except in case of the monks data set, where it improves accuracy.

6 Conclusions and Further Work

In this paper, we have considered the data mining task of predictive clustering. This is a very general task that contains many features of (and thus to a large extent generalizes over) the tasks of predictive modelling and clustering. While this task has been considered before, we have defined it both more precisely and in a more general form (i.e., to consider distances on both target and attribute variables and to consider clustering rules in addition to trees).

We have introduced the notion of clustering rules and focused on the task of learning predictive clustering rules for multi-objective prediction. The task of inducing PCRs generalizes the task of rule induction, extending it to multi-objective classification, regression and in general prediction. It also generalizes some forms of distance-based clustering and in particular itemset constrained clustering (e.g., it allows both discrete and continuous variables on the feature/attribute side, as well as the objective/target side).

Learning PCRs and predictive clustering in general can be viewed as constrained clustering, where clusters that have an explicit representation in a language of constraints are sought. At present PCR clusters are arbitrary rectangles in the attribute space, as arbitrary conjunctions of conditions are allowed in the rule antecedents. However, one can easily imagine additional language constraints being imposed on rule antecedents.

Viewing precitive clustering as constrained clustering makes it directly related to constraint-based data mining and inductive databases (IDBs). Constraint-based clustering is an under-researched topic in constraint-based data mining and the present research is a step towards rectifying this. Bringing the two most common data mining tasks closer together (as done in predictive clustering) moves us towards finding a general framework for data mining, which is also the main goal of IDBs.

We have implemented a preliminary version of a system for learning PCRs for multi-objective classification. We have also performed some preliminary experiments on several data sets. The results show that a single rule-set for MOC can be as accurate as the collection of rule-sets for individual prediction of each target. The accuracies are also comparable to those of predictive clustering trees. Experiments in varying the weight of target vs. non-target attributes in the compactness heuristic used in the search for rules show that non-zero weights for non-targets increase overall compactness and sometimes also accuracy.

Note, however, that many more experiments are necessary to evaluate the proposed paradigm and implementation. These would include experiments on additional data sets for multi-objective prediction, where classification, regression and a mixture thereof should be considered. Also, a comparison to other approaches to constrained clustering would be in order.

Other directions for further work concern further development of the PCR paradigm and its implementation. At present, our implementation only considers multi-objective classification, but can be easily extended to regression problems, and also to mixed, classification/regression problems. Currently, the heuristic guiding the search for rules does not take the number of covered examples in consideration. Consequently, construction of overly specific rules can only be prevented by means of setting the minimum number of examples covered by a rule. Adding a coverage dependent part to the heuristic would enable the induction of compact rules with sufficient coverage. Another possibility is the use of some sort of significance testing analogous to significance testing of the target variable distribution employed by CN2.

Finally, the selection of weights for calculating the distance measure (and the compactness heuristic) is an open issue. One side of this is the weighting of target vs. non-target variables. Another side is the assignment of relevance-based weights to the attributes: while this has been considered for single-objective classification, we need to extend it to multi-objective prediction.

Acknowledgements

This work was supported by the EU project *Inductive Queries for Mining Patterns and Models (IQ)*, contract 516169. Jan Struyf is a post-doctoral fellow of the Fund for Scientific Research of Flanders (FWO-Vlaanderen). Many thanks to Hendrick Blockeel for his useful comments on an earlier draft of this paper.

References

1. Blockeel, H. (1998): *Top-down induction of first order logical decision trees.* PhD thesis, Department of Computer Science, Katholieke Universiteit, Leuven.
2. Blockeel, H., De Raedt, L., and Ramon, J. (1998): Top-down induction of clustering trees. *Proceedings of the 15th International Conference on Machine Learning,* pages 55–63, Morgan Kaufmann.
3. Blockeel, H. and Struyf, J. (2002): Efficient algorithms for decision tree cross-validation, *Journal of Machine Learning Research,* 3(Dec):621–650, Microtome Publishing.
4. Clark, P. and Niblett, T. (1989): The CN2 Induction Algorithm, *Machine Learning,* 3:261–283, Kluwer.
5. Džeroski, S., Demšar, D., and Grbović, J. (2000): Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence,* 13(1): 7–17.
6. Džeroski, S., Blockeel, H., and Grbović. (2001): Predicting river water communities with logical decision trees. Presented at the Third European Ecological Modelling Conference, Zagreb, Croatia.
7. Flach, P. and Lavrač, N. (1999): Rule induction. In *Intelligent Data Analysis,* eds. Berthold, M. and Hand, D. J., pages 229–267, Springer.
8. Karalič, A. and Bratko, I. (1997): First Order Regression. *Machine Learning,* 26:147–176, Kluwer.
9. Kaufman, L. and Rousseeuw, P. J. (1990): *Finding groups in data: An introduction to cluster analysis,* John Wiley & Sons.
10. Langley, P. (1996): *Elements of Machine Learning.* Morgan Kaufman.
11. Lavrač, N., Kavšek, B., Flach, P., and Todorovski, L. (2004): Subgroup discovery with CN2-SD, *Journal of Machine Learning Research,* 5(Feb):153–188, Microtome Publishing.
12. Michalski, R. S. (1980): Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for partitioning data into conjunctive concepts. *International Journal of Policy Analysis and Information Systems,* 4:219–243.
13. Newman, D. J., Hettich, S., Blake, C. L., and Merz, C. J. (1998): *UCI Repository of machine learning databases.* University of California, Irvine, CA.
14. Quinlan, J. R. (1993): *C4.5: Programs for Machine Learning.* Morgan Kaufmann.
15. Sese, J. and Morishita, S. (2004): Itemset Classified Clustering. *Proceedings of the Eighth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04),* pages 398–409, Springer.
16. Sese, J., Kurokawa, Y., Kato, K., Monden, M., and Morishita, S. (2004) Constrained clusters of gene expression profiles with pathological features. *Bioinformatics.*
17. Struyf, J., and Dzeroski, S. (2005): Constraint based induction of multi-objective regression trees. *Proceedings of the 4th International Workshop on Knowledge Discovery in Inductive Databases (KDID 2005)*, pages 110-121.

18. Struyf, J., Dzeroski, S., Blockeel, H., and Clare, A. (2005): Hierarchical multi-classification with predictive clustering trees in functional genomics. *Proceedings of Workshop on Computational Methods in Bioinformatics as part of the 12th Portuguese Conference on Artificial Intelligence*, pages 272-283, Springer.
19. Suzuki, E., Gotoh,M., and Choki, Y. (2001): Bloomy Decision Tree for Multi-objective Classification. *Proceedings of the Fifth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01),* pages 436-447, Springer.
20. Todorovski, L., Blockeel, H., and Dzeroski, S. (2002): Ranking with predictive clustering trees. *Machine Learning: 13th European Conferende on Machine Learning, Proceedings*, pages 444-456, Springer.
21. Torgo, L. (1995): Data Fitting with Rule-based Regression. *Proceedings of the workshop on Artificial Intelligence Techniques (AIT'95),* Zizka, J. and Brazdil, P. (eds.), Brno, Czech Republic.