Incremental Approach to Classification Learning

Xenia Alexandre Naidenova
Research Centre of Military Medical Academy – Saint Petersburg, Russia

INTRODUCTION

By classification we mean partition of a given object's set into disjoint blocks or classes. We assume that objects are described by a set U of symbolic or numeric attributes and each object can have one and only one value of each attribute. Then each attribute generates, by its values, partition of a given set of objects into mutually disjoint classes the number of which is equal to the number of values of this attribute. To give a target classification of objects, we use an additional attribute KL not belonging to U. In Table 1, we have two classes: KL+ (positive objects) and KL− (negative objects).

By classification learning we mean approximation of given object classification in terms of attributes names or values of attributes (Naidenova, 2012). This approximation is reduced to extracting logical rules in the form of functional or implicative dependencies from observable datasets. These dependencies allow to distinguish between classes of given classification. For our example in Table 1, we have some rules based on implicative (ID) and functional dependencies (FD): Color_of_Hairs, Color_of_Eyes → KL (FD); if Blond, Blue, then KL = "+"; if Hazel, then KL = "−"; if Brown, Blue, then KL = "−" (IDs).

The task of classification learning based on inferring implicative rules is equivalent to the task of concept formation (Banerji, 1969, Ganter & Wille, 1999). The goal of this task is to describe/classify new objects according to description/classification of existing objects. Inferring good diagnostic (classification) tests (GDTs) is the formation of the best descriptions of a given object class KL+ against the objects not belonging to this class (KL−).

Let $M = (\cup\text{dom}(attr), attr \in U)$, where dom(attr) is the set of all values of attr. Let $X \subseteq M$ and G be the set of indices of objects considered (objects for short), $G = G+ \cup G-$, where G+ and G− the sets of positive and negative objects, respectively. Denote by d(g) the description of object $g \in G$. Let $P(X) = \{g \mid g \in G, X \subseteq d(g))$. We call P(X) the interpretation of X in the power set 2^G. If P(X) contains only positive objects and the number of these objects more than 2, then we call X a description of some positive objects and (P(X), X) a test for positive objects. Let us define a good test or good description of objects.

Definition 1: A set $X \subseteq M$ of attribute values is a good description of positive (negative) objects if and only if it is the description of these objects and no such subset $Y \subseteq M$ exists, that $P(X) \subset P(Y) \subseteq G+$ ($\subseteq G-$).

Table 1. Example of classification

Index of Example	Height	Color of Hair	Color of Eyes	KL
1	Low	Blond	Blue	+
2	Low	Brown	Blue	–
3	Tall	Brown	Hazel	–
4	Tall	Blond	Hazel	–
5	Tall	Brown	Blue	–
6	Low	Red	Blue	–
7	Tall	Red	Blue	+
8	Tall	Blond	Blue	+

DOI: 10.4018/978-1-5225-2255-3.ch017

It has been shown (Naidenova, 1992) that the problem of good tests inferring is reduced to searching for implicative dependencies in the form $X \rightarrow v$, $X \subseteq M$, $v \in dom(KL)$ for all positive (negative) objects.

The concept of good classification (diagnostic) test has firstly been introduced in (Naidenova & Polegaeva, 1986). In (Naidenova, 2012), it is considered the link between classification learning based on inferring good tests and formal concepts in the FCA.

BACKGROUND DEFINITIONS

Let $G = \{1, 2,\ldots, N\}$ be the set of objects' indices (objects, for short) and $M = \{m_1, m_2, \ldots, m_j, \ldots$ mq$\}$ be the set of attributes' values (values, for short). Each object is described by a set of values from M. The object descriptions are represented by rows of a table R the columns of which are associated with the attributes taking their values in M. Let $D(+)$ and $G(+)$ be the sets of positive object descriptions and the set of indices of these objects, respectively. Then $D(-) = D/D(+)$ and $G- = G/G+$ are the sets of negative object descriptions and indices of these objects, respectively.

The definition of good tests as a dual construction or formal concept is based on two mapping $2^G \rightarrow 2^M$, $2^M \rightarrow 2^G$ determined as follows. $A \subseteq G$, $B \subseteq M$. Denote by Bi, Bi $\subseteq M$, $i = 1,\ldots, N$ the description of object with index i. We define the relations $2^G \rightarrow 2^M$, $2^M \rightarrow 2^G$ as follows: $A' = \mathrm{val}(A) = \{$intersection of all Bi: Bi $\subseteq M$, $i \in A\}$ and $B' = \mathrm{obj}(B) = \{i: i \in G, B \subseteq Bi\}$. These mapping are the Galois's correspondences (Ore, 1944). Of course, we have $\mathrm{obj}(B) = \{$intersection of all $obj(m)$: $obj(m) \subseteq G$, $m \in B\}$. Operations val(A), obj(B) are reasoning operations (derivation operations).

We introduce two generalization operations: generalization_of(B) = B′′ = ; generalization_of(A) = A′′ =. These operations are the closure operations (Ore, 1944).

A set A is closed if A = obj(val(A)). A set B is closed, if B = val(obj(B)). For g ∈ G and m ∈ M, g′ is called object intent and m′ is called value extent.

By using the dataset in Table 1, we illustrate the derivation and generalization operations: A = {7, 8}, val(A) = {Tall, Blue}; A′′ = obj({Tall Blue}) = {5, 7, 8};

m = {Red}, obj ({m}) = {6, 7}; m′′ = val({6, 7}) = {Red, Blue};

B = {Low, Blue}, obj({B}) = {1, 2, 6}; B′′ = val({1, 2, 6}) = {Low, Blue} = B.

Classification of objects are defined as follows (Kuznetsov, 1999). Let a context K = (G, M, I) be given, where I ⊆ G × M. In addition to values of M, a target value ω ∉ M is considered. The set of objects G is partitioned into two subsets: G+ of objects with property ω (positive objects) and G− without this property (negative objects).

Then K = K+ ∪ K− ; K+ ∩ K − = ∅; K+ = (G+, M, I+); K− = (G−, M, I−); G = G+ ∪ G−; G+ ∩ G− = ∅.

Diagnostic test is defined as follows.

Definition 2: A diagnostic test for G+ is a pair (A, B) such that B ⊆ M (A = obj(B) ≠ ∅), A ⊆ G+, and ∀g, g ∈ G−: B ⊄ val(g) and B ≠ val(g).

Definition 3: A diagnostic test (A, B) such that B ⊆ M (A = obj(B) ≠ ∅) is a good test for G+ if and only if extension A* = A ∪ g, g ∉ A, g ∈ G+ implies that (A*, val(A*)) is **not a test** for G+.

Definition 4: A diagnostic test (A, B), B ⊆ M (A = obj(B) ≠ ∅) is a good maximally redundant test (GMRT) for G+ if any extension B* = B ∪ m, m ∉ B, m ∈ M implies that obj(B*) ⊆ obj(B) ((obj(B*), B*) is a test for G+ but not good).

It is important to note that if a pair (A, B) is a maximally redundant test, then A and B are closed and, consequently, this test is a formal concept in terms of the Formal Concept Analysis (the FCA)

Related Content