Classification Algorithm
Based on Pairwise Comparison of Features

E. K. Kornoushenko

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: ekorno@mail.ru
Received March 17, 2015

Abstract—We propose an alternative approach to classification that differs from known approaches in that instead of comparing the tuple of values of a test object’s features with similar tuples of features for objects in the training set, in this approach we make independent pairwise comparisons of every pair of feature values for the objects being compared. Here instead of using the notion of a “nearest neighbors” for test object, we introduce the notion of “admissible proximity” for each feature value in the test object. In this approach, we propose an alternative algorithm for classification that has a number of significant practical features. The algorithm’s quality was evaluated on sample problems taken from the well-known UCI repository and related to various aspects of human activity. The results show that the algorithm is competitive compared to known classification algorithms.

Keywords: classification, nearest neighbors, admissible proximity, class label, weighted majority, weighting features, filling of classes.

DOI: 10.1134/S000511791711011X

1. INTRODUCTION

Classification as a decision making process is one direction in the vast field of data mining that has had numerous works devoted to it by the preset time. As classification methods were created, settings of classification problems and algorithms for solving them became more and more complicated; together with quantitative data researchers increasingly use qualitative data as well. This has led to a significant extension of the set of theoretical and practical problems that can be solved (in medicine, economics, psychology, and so on), and the spectrum of classification principles has accumulated very different directions: from the simplest “nearest neighbors” principle to ordinal classification by qualitative data.

A significant place in classification literature is taken by publications based on the notion of “nearest neighbors”: a classified test object (represented as a tuple of its features) is compared with the objects in the training set, and the chosen number of nearest (according to some metric) objects for the test object form its “nearest neighborhood” (see, e.g., [1]). In the present work, we propose a modification of this approach: instead of comparing tuples of features for every pair objects we propose to consider pairwise independent comparisons of values for each feature of these objects. While the result of comparing two values of every nominal feature can only be their matching (or not matching), for compared real values of a quantitative feature we define a measure of “allowed” proximity, called “admissible proximity” of the values.1 Here instead of the notion of “nearest neighbors” for the test object we introduce the notion of “admissible close neighborhood” for each value of the feature from the tuple of features of the test object. In other words, an “admissible close neighborhood” for a given feature value contains all “admissibly

1 The choice of the “admissible proximity” value is discussed below.
close” values of this feature in tuples of features for training set objects, and these objects from
training set themselves form the “admissible neighborhood” of the test object with respect to this
feature. Further, instead of known operations on the tuples of features for objects from the “nearest
neighborhood” of the test object (weighted majority, weighting features, weighting “confidence”
coefficients—see Sections 2.2.1−2.2.3), we construct the so-called matrix of weights for the test
object, which reflects the distribution of weights of (“admissibly close” by feature values) objects
with respect to features and classes defined on objects from the training set. This approach and the
classification algorithm proposed below have a number of properties that are lacking in previously
known approaches to classification. To emphasize these properties, below in Section 2 we give a
brief survey of feature weighting procedures for objects and class labels of objects—such procedures
are used in various classification algorithms based on the “nearest neighbors” principle.

The quality of the proposed classification algorithm has been demonstrated on a number of
specific samples taken from a well-known UCI repository [2] and related to various aspects of human
activity. We note its characteristic features and possibility to apply the algorithm in practical
studies. We pay special attention to correct classification conditions for objects of “small” classes
from the test sample implemented in the algorithm, due to the fact that in practical problems
correct classification of objects from “small” classes can be of special interest.

The paper is organized as follows. Section 2 is devoted to a well-known approach based on the
notion of “nearest neighborhood”; we show a brief survey of the evolution of approaches to this
method (transition from unweighted majority to some kind of weighted majority, to reweighting
the features of objects, weighting “confidence” coefficients for the objects from the “nearest neigh-
borhood” of a test object, the use of different metrics in comparing the objects). In Section 3, we
define the stages of the proposed algorithm and show its outline. Section 4 shows the character-
istic features and differences of procedures used at a certain stage of the algorithm from similar
procedures mentioned in Section 2. In Section 5 we show an application of the proposed algorithm
to specific samples and discuss the results. The Appendix shows on a specific example that the
proposed algorithm without finding the most informative features (see Subsection 2.2.2) can yield
better results than feature weighting procedures.

2. WEIGHTING PROCEDURES IN CLASSIFICATION ALGORITHMS
THAT USE THE NOTION OF “NEAREST NEIGHBORHOOD”

2.1. Basic Notions

Initial information in the supervised classification problem considered below is a sample of
objects consisting of the training sample (TrS) of length $S$ and test sample (TsS) of length $T$.

We call elements of these samples objects. Each object is characterized by a collection of features.

A feature (attribute) of an object is a certain characteristic of the object used for its classification.

Each object in the original sample has $m$ of the same features in both TrS and TsS. Features can
be quantitative (continuous) with real values from the corresponding intervals and qualitative with
finite sets of (verbal or numeric) values on the corresponding scales. A tuple of features for an
object is the collection of values of the features characterizing this object. In what follows we will
treat “object” and “tuple of features of an object” as synonyms for simplicity.

We assume that on the set of objects in the original sample there is defined a partition $\pi$ into
classes, and each object from TrS and TsS belongs to some class of this partition. We denote by $\pi_{\text{tr}}$
and $\pi_{\text{test}}$ subpartitions of partition $\pi$ such that blocks of partition $\pi_{\text{tr}}$ contain only objects from TrS
and blocks of partition $\pi_{\text{test}}$ contain only objects from TsS. The index of a block from partition $\pi$
which a given object belongs to is called the label of this object. Here each block of partition $\pi$
may contain objects from both TrS and TsS (but cannot contain only objects from TsS). If some
block contains objects from both TrS and TsS, labels of such objects coincide, and collections of