Musk (Version 2) Data Set

Download: Data Folder, Data Set Description

Abstract: The goal is to learn to predict whether new molecules will be musks or non-musks

Data Set Characteristics:	Multivariate	**Number of Instances:**	6598	**Area:**	Physical
Attribute Characteristics:	Integer	**Number of Attributes:**	168	**Date Donated**	1994-09-12
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	82526

Source:

Creators:

AI Group at Arris Pharmaceutical Corporation
contact: David Chapman or Ajay Jain
Arris Pharmaceutical Corporation
385 Oyster Point Blvd.
South San Francisco, CA 94080
415-737-8600
zvona '@' arris.com, jain '@' arris.com

Donor:

Tom Dietterich
Department of Computer Science
Oregon State University
Corvallis, OR 97331
503-737-5559
tgd '@' cs.orst.edu

Data Set Information:

This dataset describes a set of 102 molecules of which 39 are judged by human experts to be musks and the remaining 63 molecules are judged to be non-musks. The goal is to learn to predict whether new molecules will be musks or non-musks. However, the 166 features that describe these molecules depend upon the exact shape, or conformation, of the molecule. Because bonds can rotate, a single molecule can adopt many different shapes. To generate this data set, all the low-energy conformations of the molecules were generated to produce 6,598 conformations. Then, a feature vector was extracted that describes each conformation.

This many-to-one relationship between feature vectors and molecules is called the "multiple instance problem". When learning a classifier for this data, the classifier should classify a molecule as "musk" if ANY of its conformations is classified as a musk. A molecule should be classified as "non-musk" if NONE of its conformations is classified as a musk.

Attribute Information:

molecule_name: Symbolic name of each molecule. Musks have names such as MUSK-188. Non-musks have names such as NON-MUSK-jp13.
conformation_name: Symbolic name of each conformation. These have the format MOL_ISO+CONF, where MOL is the molecule number, ISO is the stereoisomer number (usually 1), and CONF is the conformation number.
f1 through f162: These are "distance features" along rays (see paper cited above). The distances are measured in hundredths of Angstroms. The distances may be negative or positive, since they are actually measured relative to an origin placed along each ray. The origin was defined by a "consensus musk" surface that is no longer used. Hence, any experiments with the data should treat these feature values as lying on an arbitrary continuous scale. In particular, the algorithm should not make any use of the zero point or the sign of each feature value.
f163: This is the distance of the oxygen atom in the molecule to a designated point in 3-space. This is also called OXY-DIS.
f164: OXY-X: X-displacement from the designated point.
f165: OXY-Y: Y-displacement from the designated point.
f166: OXY-Z: Z-displacement from the designated point.
class: 0 => non-musk, 1 => musk

Please note that the molecule_name and conformation_name attributes should not be used to predict the class.

Relevant Papers:

Dietterich, T. G., Jain, A., Lathrop, R., Lozano-Perez, T. (1994). A comparison of dynamic reposing and tangent distance for drug activity prediction. Advances in Neural Information Processing Systems, 6. San Mateo, CA: Morgan Kaufmann. 216--223.
[Web Link]

Jain, A. N., Dietterich, T. G., Lathrop, R. H., Chapman, D., Critchlow, R. E., Bauer, B. E., Webster, T. A., Lozano-Perez, T. Compass: A shape-based machine learning tool for drug design. Computer-Aided Molecular Design.
[Web Link]

Dietterich, T. G., Lathrop, R. H., Lozano-Perez, T. Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence.
[Web Link]