Computers in Biology and Medicine 41 (2011) 115–122

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A new dataset evaluation method based on category overlap

Sejong Oh*

Department of Nanobiomedical Science, Dankook University, Cheonan 330-714, Republic of Korea

ARTICLE INFO

Article history:
Received 17 February 2010
Accepted 22 December 2010

Keywords:
Feature
Feature selection
R-value
Dataset
Classification
Machine learning algorithm

ABSTRACT

The quality of dataset has a profound effect on classification accuracy, and there is a clear need for some method to evaluate this quality. In this paper, we propose a new dataset evaluation method using the *R-value* measure. This proposed method is based on the ratio of overlapping areas among categories in a dataset. A high *R-value* for a dataset indicates that the dataset contains wide overlapping areas among its categories, and classification accuracy on the dataset may become low. We can use the *R-value* measure to understand the characteristics of a dataset, the feature selection process, and the proper design of new classifiers.

1. Introduction

Classification is one of the primary themes in computational biology. For example, researchers have a need to classify or predict the protein subcellular locations of new and unknown proteins. In the service of this task, they derive a specific *dataset*, and then train *classifiers* using known classified protein data. The accuracy of training and classification depends on the quality of the dataset. *Feature selection* [22,23,28–31] is the problem of selecting a small subset of features, which is, ideally, necessary and sufficient to describe the target concept [1]. The objective of feature selection is to obtain a feature space characterized by (1) low dimensionality, (2) retention of sufficient information, (3) enhancement of separability in feature space for examples in different categories via the removal of effects resulting from noisy features, and (4) the comparability of features among examples in the same category [2]. As a result of feature selection, we can take a dataset for a classification task.

With regard to the clarity of dataset-associated terms, we would like to define the terms using the data in Table 1. Table 1 shows the hypothetical experiment data for research disease A. The data was abstracted from 200 people, and the row in the table is designated as *instance*. The subjects can be classified into two *categories*: 'patient' and 'normal'. A more general term for 'category' is 'class', but we use 'category' for the sake of clarity. Table 1 contains four data columns—'height', 'weight', 'running hour', and 'working hour'; we call these *features* (or *attributes*). Columns 'No' and 'etc' can be called attributes, but we exclude the columns containing auxiliary or category information from the features. We can derive a dataset that contains at least one of the features in Table 1 for the training and testing of our classifier. For example, we can derive instances of feature {height} for a dataset. We can also derive multiple features, such as {height, weight}, {weight, running hour, working hour}, and {height, weight, running hour, working hour} for a dataset. If an experimental data has m features, there will exist (2^m-1) datasets. A *dimension* of a dataset is defined by the number of features a dataset includes.

In order to select optimal features from a given set of features, we need to evaluate each derived dataset via evaluation functions (or measures). The majority of evaluation functions use a scoring scheme. Let us suppose that D_1 and D_2 are datasets and $E(x)$ is an evaluation function, and thus $E(D_1)$ and $E(D_2)$ generate some scores. If $E(D_1) > E(D_2)$, we can expect that dataset D_1 will yield better training/testing accuracy than D_2. Dash and Liu [3] grouped evaluation functions into five categories—*distance measures*, *information measures*, *dependency measures*, *consistency measures*, and *classifier error rate measures.* Distance measures [24] are the most popular, and include separability, divergence, and discrimination measures. The Euclidean distance measure is a typical distance measure. In Section 2, we summarize more distance measures.

In this paper, we propose a new method for dataset evaluation based on category overlap. It belongs to the distance measure group. The method proposed herein is based on the following assumption: if dataset D_1 makes more separable categories than dataset D_2, then D_1 yields better classification accuracy than D_2. We also believe that separability is strongly related to category overlap. Fig. 1 shows an example of this. There are six datasets with different degrees of overlap. Each dataset contains two-dimensional features and three

* Tel.: +82 41 550 3484; fax: +82 41 550 1149.
E-mail address: sejongoh@dankook.ac.kr

0010-4825/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2010.12.006

categories. The order of overlap degree is as follows: $D_1 < D_2 < D_3 < D_4 < D_5 < D_6$. D_4 and D_5 are almost identical, but D_5 has a slightly larger overlapping area than D_4. We may expect that the accuracy of classification from the datasets would be as follows: $D_1 > D_2 > D_3 > D_4 > D_5 > D_6$. Fig. 2 shows the experimental results of well-known classifiers on the features in Fig. 1. We experiment with Naïve Bayes (NB) [21], *K*-Nearest Neighbors (KNN) [27], Artificial Neural Network (ANN), and Support Vector Machine (SVM) classifiers [26]. The results tell us that our expectations are reasonable.

The separability of a dataset is strongly related to the degree of overlap among categories in the dataset. If we measure the degree of overlap, we can determine which dataset is better than others. We can also develop a feature selection algorithm using the measures. This is the motivation of the work, and we develop an *R-value* measure as a new dataset evaluation method. If *R-value*($dataset_1$) is higher than *R-value*($dataset_2$), then $dataset_1$ is considered to have a broader area of overlap than $dataset_2$, and $dataset_2$ yields higher classification accuracy than $dataset_1$.

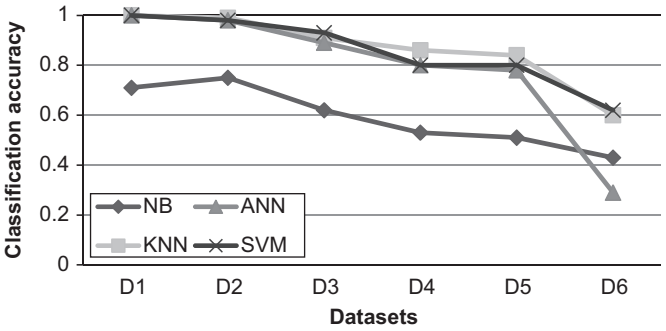
The remainder of this paper is structured as follows. Section 2 summarizes several dataset evaluation methods, and shows that the methods cannot adequately capture the separability of datasets. Section 3 describes the proposed dataset evaluation method using *R-value*. Section 4 describes the results of experiments concerning *R-value*. We compare *R-value* with other evaluation methods in terms of classification accuracy and quality. Section 5 summarizes the advantages of the *R-value* method, and the conclusions of this paper are provided in Section 6.

Table 1
Experimental data for disease A.

No	Height	Weight	Running hour	Working hour	etc
1	0.41	0.36	0.27	0.65	Patient
2	0.23	0.37	0.34	0.68	Patient
3	0.38	0.38	0.46	0.95	Patient
4	0.45	0.31	0.37	0.75	Patient
5	0.37	0.45	0.48	0.75	Patient
…	…	…	…	…	…
195	0.89	0.56	0.81	0.56	Normal
196	0.65	0.57	0.81	0.43	Normal
197	0.75	0.67	0.76	0.35	Normal
198	0.46	0.48	0.65	0.42	Normal
199	0.89	0.69	0.78	0.23	Normal
200	0.78	0.81	0.88	0.26	Normal

2. Study of dataset evaluation methods

In this section, we summarize several distance measures for dataset evaluation. These measures are designed for feature evaluation, but we expend and adopt them for dataset evaluation. The goal of distance measurement is to calculate the distances among categories in a dataset. Let D be a distance function and



Classifier	D_1	D_2	D_3	D_4	D_5	D_6
NB	0.71	0.75	0.62	0.53	0.51	0.43
KNN	1.0	0.99	0.91	0.86	0.84	0.6
ANN	1.0	0.98	0.89	0.80	0.78	0.29
SVM	1.0	0.98	0.93	0.80	0.80	0.62

Fig. 2. Comparison of classification accuracies for datasets in Fig. 1.

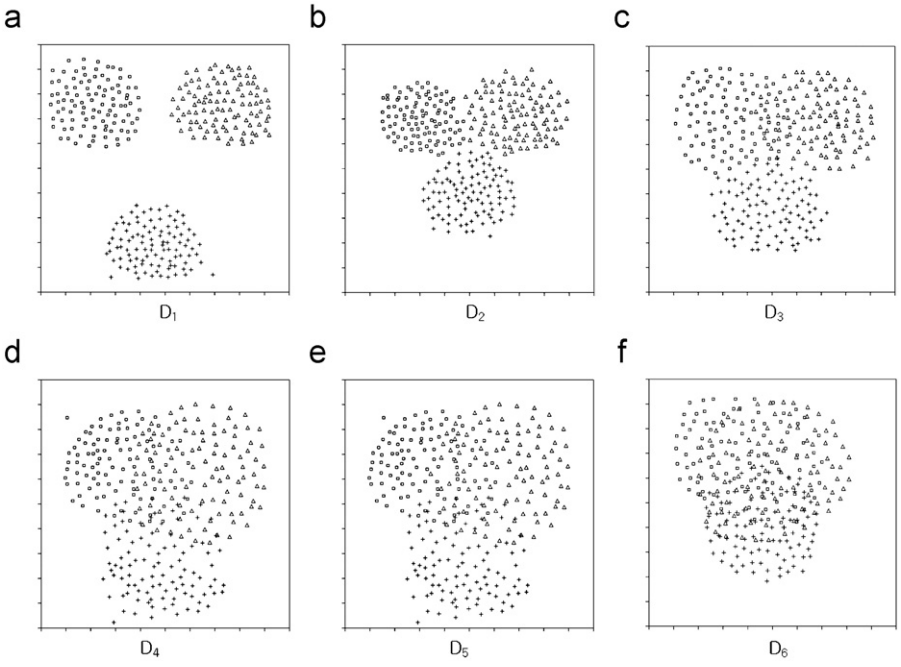


Fig. 1. Six datasets that have different overlap degrees.