SBL-MMRC: A k-Nearest-Neighbor Method with Multi-Metric Computing of Reference Vectors

Karol Grudziński

Department of Physics, Academy of Bydgoszcz
Plac Weyssenhoffa 11, 85-072 Bydgoszcz, Poland
E-mail: karol.grudzinski@wp.pl

and

Institute of Applied Informatics
University of Economy
Garbary 2, 85-229 Bydgoszcz, Poland

Abstract. Multi-Metric Reference Calculation Models (SBL-MMRC), which are introduced in this paper, are k-Neareast Neighbors method based classifiers that involve computing reference instances for each query case with different (multiple) similarity metrics. These methods are lazy learning models. In this paper three variants of the SBL-MMRC algorithm are compared to classical metric optimization on four real-world datasets. In the all experiments at least one of the SBL-MMRC models outperformed slightly classical metric optimization method based on iterating over the metrices. In most SBL-MMRC experiments the variance obtained on the test set has been lower than the one attained by classical metric optimization.

keywords: k-Nearest-Neighbor metric optimization, similarity-based methods, classification of data.

1 Introduction.

The k-Nearest-Neighbors method (k-NN) [] is a very popular data classification model which is very often used as a reference method by many researchers. The classification accuracy obtained with the basic model on most of the data can be often significantly increased by numerous extensions and modifications of this algorithm. However, the entire power of the k-NN method, the familiarity with its extensions and the knowldedge about its weak sides, is known to a rather narrow subgroup of pattern recognition researchers. It is hard to find a good public domain k-NN program, especially such one that is equipped with a grafical user interface and having built-in less common extensions. It is the aim of the SBL system [] (so far existing only in the experimental protoype version), to fill this gap of user-friendly, public-domain k-NN programs.

In order to classify an unseen (query) vector q with the k-Nearest Neighbors method one computes a similarity function between q and the training vectors in a database. In the basic variant of the k-NN algorithm the predicted class of q is estimated by taking the majority class of the q's k-nearest neighobrs. Various similarity functions may be applied leading to different classifiction results.

In this paper a familly of the three methods which involve computing reference instances for each query case with different similarity metrics is introduced. This group of algorithms is called SBL-MMRC (which stands for "k-Nearest-Neighbor Method with Multi-Metric Computing of Reference Vectors"). The described methods are lazy-learning k-NN extensions and can be the alternative to classical metric optimization. The studied algorithms can be used as built-in models into other k-NN methods involving learning such as the optimization of the number of nearest neighbors or in the attribute selection.

2 The Metric-Optimization Models

In this section the four methods used in our numerical experiments are described. First the classical metric optimization method is presented in a form as it is implemented in the SBL system. Next the similarity metrics that are used in computing reference cases in all the numerical experiments are listed. After that each of the three algorithms contributing to the SBL-MMRC family is described.

2.1 The Classical Metric-Optimization Method

Selection of similarity-function very strongly influences the results that are obtained with the k-NN method. The metric optimization algorithm which is used in the SBL system is very simple: the cross-validation procedure is performed on training data and the subsequent metrics from the optimization pool of available functions are taken each time to compute the cross-validation training accuracy. The similarity function corresponding to the highest training accuracy is selected as the most promising and with the similarity function chosen in this way the unseen samples are classified.

2.2 The Similarity Metrics

2.3 The SBL-MMRC Models

The learning was conducted with the leave-one-out cross validation whilst the unseen data has been classified with 10-fold stratified cross-validation. It is a liitle bit unfortunate that different cross-validation test has been applied to training and to test data as training results indicate that usually classical metric selection should be performed on unseen data.

References

Table 1. Results for the 10-fold CV test on appendicitis data.

System	k	Train %	Test %
SBL-MMRC-Smart-FT	$k = 1$	83.6 ± 0.1	83.2 ± 2.8
SBL-MMRC-Smart,	$k = 1$	82.0 ± 0.0	82.3 ± 1.3
SBL-Metric-Selection,	$k = 1$	85.6 ± 0.1	80.2 ± 3.5
SBL-MMRC-Basic,	$k = 1$	74.9 ± 0.0	76.4 ± 1.9
SBL-MMRC-Smart-FT,	$k = 10$	87.3 ± 0.0	87.3 ± 0.2
SBL-MMRC-Smart,	$k = 10$	86.4 ± 0.0	86.2 ± 0.6
SBL-MMRC-Basic,	$k = 10$	85.7 ± 0.0	85.8 ± 0.0
SBL-Metric-Selection,	$k = 10$	88.4 ± 0.0	84.7 ± 2.5

Table 2. Results for the 10-fold CV test on Cleveland heart data.

System	k	Train %	Test %
SBL-MMRC-Smart-FT,	$k = 1$	74.6 ± 0.0	75.4 ± 1.4
SBL-Metric-Selection,	$k = 1$	78.5 ± 0.0	75.0 ± 2.0
SBL-MMRC-Smart,	$k = 1$	74.7 ± 0.0	74.9 ± 1.0
SBL-MMRC-Basic,	$k = 1$	64.1 ± 0.0	64.6 ± 1.2
SBL-MMRC-Smart-FT,	$k = 10$	81.4 ± 0.0	81.8 ± 0.2
SBL-MMRC-Smart,	$k = 10$	81.4 ± 0.0	81.6 ± 1.3
SBL-Metric-Selection,	$k = 10$	83.4 ± 0.1	80.8 ± 0.9
SBL-MMRC-Basic,	$k = 10$	80.7 ± 0.1	80.8 ± 1.4

Table 3. Results for the 10-fold CV test on iris data.

System	k	Train %	Test %
SBL-MMRC-Basic,	$k = 1$	96.6 ± 0.0	96.9 ± 0.3
SBL-Metric-Selection,	$k = 1$	96.9 ± 0.0	95.9 ± 0.3
SBL-MMRC-Smart,	$k = 1$	95.3 ± 0.0	95.3 ± 0.1
SBL-MMRC-Smart-FT,	$k = 1$	92.5 ± 0.0	95.2 ± 0.3
SBL-MMRC-Smart,	$k = 10$	97.0 ± 0.0	96.9 ± 0.3
SBL-MMRC-Basic,	$k = 10$	96.8 ± 0.0	96.8 ± 0.2
SBL-Metric-Selection,	$k = 10$	98.0 ± 0.0	96.7 ± 0.7
SBL-MMRC-Smart-FT,	$k = 10$	84.2 ± 0.2	96.0 ± 0.0

Table 4. Results for the 10-fold CV test on hepatitis data.

System	k	**Train %**	**Test %**
SBL-MMRC-Smart ,	$k = 1$	79.3 ± 0.0	79.7 ± 1.0
SBL-Metric-Selection,	$k = 1$	82.1 ± 0.0	77.8 ± 2.0
SBL-MMRC-Smart-FT,	$k = 1$	75.0 ± 0.0	76.3 ± 6.5
SBL-MMRC-Basic,	$k = 1$	72.1 ± 0.0	73.3 ± 1.4
SBL-MMRC-Smart,	$k = 10$	86.1 ± 0.1	87.0 ± 0.9
SBL-Metric-Selection,	$k = 10$	87.6 ± 0.0	85.0 ± 3.6
SBL-MMRC-Smart-FT,	$k = 10$	83.2 ± 0.0	82.8 ± 2.9
SBL-MMRC-Basic,	$k = 10$	79.9 ± 0.0	80.9 ± 1.1