Pattern Recognition Letters 49 (2014) 62–68

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Local value difference metric ☆

Chaoqun Li [a,*], Liangxiao Jiang [b], Hongwei Li [a]

[a] Department of Mathematics, China University of Geosciences, Wuhan 430074, China
[b] Department of Computer Science, China University of Geosciences, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 16 March 2014
Available online 9 July 2014

Keywords:
Value difference metric
Distance metric learning
Distance-based learning algorithms
Local learning
Decision tree learning

ABSTRACT

Value difference metric (VDM) is one of the widely used distance functions designed to work with nominal attributes. Research has indicated that the definition of VDM follows naturally from a simple probabilistic model called a naive Bayes (NB). NB assumes that all the attributes are independent given the class. To further improve the performance of NB, several techniques have been proposed. Among these, an effective technique is local learning. Because VDM has a close relationship with NB, in this paper, we propose a local learning method for VDM. The improved distance function is called local value difference metric (LVDM). When LVDM computes the distance between a test instance and each training instance, the conditional probabilities in VDM are estimated by counting from the neighborhood of the test instance only instead of from all the training data. A modified decision tree algorithm is proposed to determine the neighborhood of the test instance. The experimental results on 43 datasets downloaded from the University of California at Irvine (UCI) show that the proposed LVDM significantly outperforms VDM in terms of the class-probability estimation performance of distance-based learning algorithms.

1. Introduction

Among the many machine learning algorithms, distance-based learning algorithms such as k-nearest neighbor (KNN), k-nearest neighbor with distance weighting (KNNDW) [1–3], and locally weighted learning algorithms [4–6] generally exhibit good performance. This is because these algorithms can construct a different approximation to the target function for each distinct test instance. These algorithms first use a distance function to determine how close a test instance is to each stored training instance. They then build a local model in the neighborhood of the test instance. Finally, the test instance is predicted using the built local model. The success of distance-based learning algorithms always depends on an effective distance function. In this paper, we focus our attention on the distance computation for nominal attributes only.

Of the numerous distance functions designed for nominal attributes, overlapping metric (OM) [1] is the simplest and is widely used. Let x be a training instance represented by an attribute vector $\langle a_1(x), a_2(x), \ldots, a_m(x)\rangle$ and a class label c_j $(j = 1, 2, \ldots, t)$, where m is the number of attributes, $a_i(x)$ is the ith attribute value of x, and t is the number of classes. Then, OM defines the distance between x and y as:

$$d(x,y) = \sum_{i=1}^{m} \delta(a_i(x), a_i(y)), \tag{1}$$

where

$$\delta(a_i(x), a_i(y)) = \begin{cases} 0, & a_i(x) = a_i(y), \\ 1, & a_i(x) \neq a_i(y). \end{cases} \tag{2}$$

According to the observation by Wilson and Martinez [7], OM fails to exploit the additional information provided by the nominal attribute values that can aid in generalization. In order to find reasonable distance between each pair of instances with nominal attributes only, value difference metric (VDM) [8] was designed. An unweighted version of VDM defines the distance between x and y as:

$$d(x,y) = \sum_{i=1}^{m} \sum_{j=1}^{t} |P(c_j|a_i(x)) - P(c_j|a_i(y))|, \tag{3}$$

where $P(c_j|a_i(x))$ is the conditional probability that the class label of x is c_j given that the ith attribute has the value $a_i(x)$.

Kasif et al. [9] showed that the definition of VDM follows naturally from a simple probabilistic model. The simple probabilistic model is a naive Bayes (NB). NB is one of the most popular data mining algorithms [10]. It has shown surprisingly good classification performance in many real-world applications despite its unrealistic attribute-independence assumption. To further improve the performance of NB, many techniques have been proposed. One of

☆ This paper has been recommended for acceptance by J. Laaksonen.

* Corresponding author. Tel./fax: +86 27 67883091.
E-mail address: chqli@cug.edu.cn (C. Li).

http://dx.doi.org/10.1016/j.patrec.2014.06.014
0167-8655/© 2014 Elsevier B.V. All rights reserved.

these techniques is local learning [4–6,11]. These proposals for local learning show that the attribute-independence assumption required by NB is likely to be true in the neighborhood of the test instance. Thus, local learning may mitigate the effects of attribute dependencies that exist in the data as a whole.

Because VDM follows naturally from NB, VDM also assumes that all the attributes are independent given the class. Inspired by local naive Bayesian models, this paper attempts to use a local learning technique to improve VDM. The improved distance function is called local value difference metric (LVDM). When called upon to compute the distance between a test instance y and each training instance x, VDM estimates the conditional probabilities $P(c_j|a_i(x))$ and $P(c_j|a_i(y))$ by counting from all the training instances. Conversely, LVDM estimates these conditional probabilities from the neighborhood of the test instance only. The neighborhood of the test instance is determined using a decision tree learning algorithm. However, traditional decision tree learning algorithms such as ID3 [12] and C4.5 [13] cannot be directly applied to this task. The induction process of the decision tree in LVDM uses a distance-based attribute selection measure [14] and an approach that restricts the growth of the tree sooner to build a smaller tree to ensure that there are sufficient instances in each leaf node to provide a reliable probability estimation for VDM.

To upwardly scale the classification performance of distance-based learning algorithms, several improved versions of VDM have been proposed. These include modified value difference metric (MVDM) [15], interpolated value difference metric (IVDM) [7], one-dependence value difference metric (ODVDM) [16], and augmented value difference measure (AVDM) [17]. However, in many real-world applications, classification alone is not always adequate [18]. For example [19], in cost-sensitive learning, the class membership probabilities are used to minimize the conditional risk [20]. Thus, accurate class-probability estimation is often required instead of classification in many real-world data-mining applications. For the task of class-probability estimation, the conditional log likelihood (CLL) [21,22] instead of the classification accuracy, is a more reasonable performance measure for the built classifiers. In a paper by Hall [23], the mean root relative squared error (RRSE) of the probability estimates is used as a performance measure to evaluate the probability-based classifiers. Thus, in this paper, we also use CLL and RRSE to evaluate the distance-based classifiers. The experimental results in Section 4 validate the effectiveness of the proposed LVDM.

The remainder of this paper is organized as follows. Section 2 reviews research on local learning. In Section 3, we propose the local value difference metric (LVDM). Section 4 reports on the empirical study. In Section 5, we draw our conclusions.

2. Related work

When a test instance y is to be classified, naive Bayes (NB) uses Eq. (4) to estimate each class membership probability $P(c_j|y)$, and then y is classified to the class c_j if the probability $P(c_j|y)$ is maximal.

$$P(c_j|y) = \frac{P(c_j)\prod_{i=1}^{m} P(a_i(y)|c_j)}{\sum_{j=1}^{t} P(c_j)\prod_{i=1}^{m} P(a_i(y)|c_j)}. \tag{4}$$

Although the attribute-independence assumption required by NB rarely holds in real-world applications, NB does show surprising classification performance. To further improve the performance of NB, researchers have proposed various improved techniques. Among these, an effective technique is local learning. Two typical representatives are naive Bayes tree (NBTree) [11] and locally weighted naive Bayes (LWNB) [5]. NBTree first builds a decision tree over the entire instance space and subsequently builds an NB classifier on each leaf node. Thus, each leaf node of an NBTree is an NB classifier. When called upon to classify a test instance y, NBTree sorts the test instance down the tree from the root to a leaf node. It then uses the NB classifier built on the leaf node to classify the test instance. NBTree is a hybrid algorithm combining decision tree learning and Bayesian learning. LWNB first finds the k-nearest neighbors of the test instance y and then constructs a local NB in the neighborhood of the test instance. The training instances in the neighborhood are weighted according their distance values to the test instance; the greater weights are given to the closer neighbors. LWNB is a hybrid algorithm combining instance-based learning and Bayesian learning. Jiang et al. [6] also proposed a local learning approach to improve the performance of naive Bayes text classifiers.

The basic idea underlying these local naive Bayesian models is that the attribute-independence assumption required by NB is more likely to be held in a local or neighborhood of the training instances than over the entire instance space. The difference is that NBTree uses a decision tree algorithm to determine the neighborhood of a test instance while LWNB uses the k-nearest neighbor algorithm (KNN) for the same.

According to Ref. [9], VDM transforms each attribute value $a_i(x)$ to a discrete probability distribution $\langle P(c_1|a_i(x)), P(c_2|a_i(x)), \ldots, P(c_t|a_i(x))\rangle$. This implies that $\langle a_1(x), a_2(x), \ldots, a_m(x)\rangle$ can be transformed into a tm-dimensional space. They call the transformation of nominal values into probability distributions a memory-based reasoning (MBR) transform. The transformation performed by VDM is induced by a simple NB model. From Eqs. (3) and (4), we can determine that VDM must estimate $P(c_j|a_i(x))$ for each class c_j and each attribute value $a_i(x)$, and NB must estimate $P(a_i(x)|c_j)$ for each class c_j and each attribute value $a_i(x)$. Thus, VDM computes the same statistics as NB.

Because the definition of VDM follows naturally from naive Bayes (NB), we would like to determine whether those techniques proposed to improve NB would be effective to improve VDM. Inspired by NBTree and LWNB, we also use the local learning technique to improve VDM in this paper. We call this improved distance function local value difference metric (LVDM). Section 3 describes the learning process of LVDM in detail.

3. Local value difference metric

In order to learn local value difference metric (LVDM), we must find the neighborhood of a test instance. Both decision tree learning algorithms and the k-nearest neighbor algorithm (KNN) can be used for this purpose. However, KNN is itself a type of distance-based learning algorithm. Moreover, VDM is generally applied to distance-based learning algorithms, which are usually lazy-learning algorithms and have high computational costs. Therefore, decision tree learning algorithms are a better choice for LVDM.

A fundamental issue in building decision trees is the determining the method to define the attribute selection measure at each non-terminal node of the tree. In existing studies, numerous measures are defined [12–14,24,25]. Among these, the information-gain measure and gain-ratio measure [12,13] are two measures that are widely used. However, in this paper, we choose the distance-based attribute selection measure [14]. A paper by Mantaras [14] shows that the accuracy of the decision tree produced by the distance-based attribute selection measure is comparative to the decision tree produced by the information-gain and gain-ratio measures. Moreover, the distance-based attribute selection measure produces smaller trees than the information-gain and gain-ratio measures, particularly in the case of datasets whose attributes have very different numbers of values. Smaller trees mean that there are more instances in each leaf node. This is helpful for the probability estimation of VDM.

First, we rewrite the definition of the distance-based attribute selection measure [14]. Let D be a dataset, P_C is the partition $\{C_1, C_2, \ldots, C_t\}$ of D in its t classes, and P_{A_i} is the partition $\{D_1, D_2, \ldots, D_n\}$ generated by the n possible values of attribute A_i. Let $I(P_C)$ and $I(P_{A_i})$, respectively, be the average information of partition P_C and P_{A_i}. $I(P_C)$ is defined as:

$$I(P_C) = -\sum_{j=1}^{t} P_j \log_2 P_j. \tag{5}$$

Similarly, $I(P_{A_i})$ is defined as:

$$I(P_{A_i}) = -\sum_{i=1}^{n} P_i \log_2 P_i. \tag{6}$$

The distance between two partitions P_C and P_{A_i} is $d(P_C, P_{A_i})$, defined as:

$$d(P_C, P_{A_i}) = I(P_C|P_{A_i}) + I(P_{A_i}|P_C), \tag{7}$$

where $I(P_C|P_{A_i}) = I(P_C \bigcap P_{A_i}) - I(P_{A_i})$, and $I(P_C \bigcap P_{A_i})$ is defined as:

$$I(P_C \bigcap P_{A_i}) = -\sum_{i=1}^{n} \sum_{j=1}^{t} P_{ij} \log_2 P_{ij}. \tag{8}$$

Based on the distance $d(P_C, P_{A_i})$, the normalized distance-based attribute selection measure is defined as:

$$d_N(P_C, P_{A_i}) = \frac{d(P_C, P_{A_i})}{I(P_C \bigcap P_{A_i})}, \tag{9}$$

where $d_N(P_C, P_{A_i})$ is a distance in $[0, 1]$. Namely, when the attribute P_{A_i} perfectly partitions the dataset D, $d_N(P_C, P_{A_i})$ achieves its minimum value 0.

The second issue in building decision trees is deciding how to control the size of the tree. The conditional probabilities in VDM are estimated directly by counting from all training instances; LVDM estimates the conditional probabilities by counting from the neighborhood of the test instance. The neighborhood of the test instance is generally defined as the leaf node into which the test instance falls. Therefore, the number of instances in each leaf node cannot be too small. Fortunately, for discrete data, the estimates tend to stabilize quickly and more data does not change the underlying mode significantly [11]. The experimental results in the LWNB paper [5] show that LWNB is relatively insensitive to the size of the neighborhood of the test instance, provided that it is not extremely small. In LVDM, let q be the number of instances in each node; q is a predetermined parameter. When the number

Table 1
CLL comparisons for VDM versus OM and LVDM.

Dataset	VDM	OM	LVDM
anneal	−94.00 ± 2.29	−132.45 ± 2.71●	−87.92 ± 3.17○
anneal.ORIG	−93.95 ± 3.71	−104.45 ± 2.61●	−94.79 ± 4.35
audiology	−110.96 ± 2.86	−117.15 ± 2.44●	−109.35 ± 3.53○
autos	−55.67 ± 2.20	−59.96 ± 1.86●	−54.18 ± 2.62○
balance-scale	−67.68 ± 5.03	−67.53 ± 4.80	−67.67 ± 5.01
breast-cancer	−33.14 ± 2.90	−32.22 ± 1.33	−33.40 ± 2.83
breast-w	−21.44 ± 2.50	−18.75 ± 2.91○	−17.92 ± 3.34○
car-evaluation	−115.60 ± 4.62	−205.74 ± 2.48●	−132.64 ± 4.22●
cardiotocography-2	−161.25 ± 6.17	−230.64 ± 5.73●	−147.58 ± 7.41○
colic	−33.06 ± 2.86	−41.26 ± 1.26●	−30.90 ± 3.33○
colic.ORIG	−35.91 ± 3.64	−45.00 ± 0.72●	−36.74 ± 3.58
connectionist-vowel	−150.66 ± 3.57	−154.73 ± 3.21●	−145.93 ± 3.72○
credit-a	−50.04 ± 3.21	−59.75 ± 2.75●	−49.50 ± 4.09
credit-g	−106.25 ± 3.99	−113.58 ± 1.89●	−109.06 ± 5.87
diabetes	−74.21 ± 6.23	−74.25 ± 4.36	−73.65 ± 6.19
glass	−41.88 ± 3.54	−41.07 ± 3.28○	−41.93 ± 3.03
heart-c	−46.32 ± 3.92	−48.79 ± 2.79●	−44.31 ± 4.12○
heart-h	−34.67 ± 5.42	−33.45 ± 4.59	−34.06 ± 5.03
heart-statlog	−22.24 ± 2.22	−21.61 ± 1.60	−22.19 ± 2.91
hepatitis	−10.98 ± 1.79	−12.60 ± 1.27●	−11.02 ± 2.16
ionosphere	−29.69 ± 1.21	−33.78 ± 1.00●	−21.84 ± 1.26○
iris	−6.87 ± 1.77	−6.53 ± 1.64	−6.89 ± 2.03
kr-vs-kp	−103.92 ± 5.42	−222.40 ± 4.09●	−79.45 ± 3.49○
labor	−3.59 ± 0.67	−4.15 ± 0.57●	−3.47 ± 0.81
letter	−8543.55 ± 36.40	−9470.64 ± 32.66●	−8169.60 ± 37.31○
libras-movement	−174.54 ± 2.35	−174.83 ± 2.28●	−169.77 ± 2.62○
lymphography	−24.43 ± 1.79	−32.02 ± 1.01●	−21.60 ± 2.08○
monks-problems-1	−34.22 ± 11.50	−44.94 ± 1.47	−38.89 ± 10.23
monks-problems-2	−58.34 ± 13.58	−72.69 ± 2.64	−61.20 ± 5.22
mushroom	−149.60 ± 0.60	−179.78 ± 1.97●	−139.88 ± 1.05○
parkinsons	−12.69 ± 1.42	−13.94 ± 1.24●	−10.74 ± 2.43○
primary-tumor	−173.11 ± 2.80	−175.44 ± 2.64●	−171.48 ± 3.25○
segment	−544.48 ± 6.05	−598.68 ± 5.40●	−421.59 ± 10.48○
sick	−86.17 ± 5.84	−95.94 ± 4.30●	−87.13 ± 4.91
sonar	−19.91 ± 1.58	−20.79 ± 1.26●	−17.06 ± 2.79○
soybean	−273.16 ± 4.95	−301.55 ± 3.97●	−231.84 ± 5.06○
splice	−584.21 ± 4.27	−665.31 ± 3.58●	−469.20 ± 6.73○
thyroid-disease	−113.10 ± 5.20	−129.91 ± 6.96●	−118.61 ± 7.36
vehicle	−143.83 ± 5.12	−150.56 ± 4.60●	−131.59 ± 5.45○
vote	−20.89 ± 2.51	−22.12 ± 2.42●	−15.80 ± 3.10○
vowel	−278.50 ± 5.28	−318.10 ± 3.93●	−279.73 ± 4.93
waveform-5000	−840.44 ± 4.05	−1004.09 ± 1.67●	−739.38 ± 7.37○
zoo	−16.86 ± 1.42	−17.61 ± 1.21●	−16.71 ± 1.39
W/T/L	–	2/8/33	23/19/1

○, ● Statistically significant upgradation or degradation.

of instances attached to a node is larger than q, the decision tree continues to grow and the instances falling into this node are partitioned recursively. However, when a test instance is sorted down the tree from the root node to a leaf node, the number of instances attached to this leaf node can be less than q. In this situation, the probability estimates are poor. To overcome this difficulty, when the number of instances in the leaf node is less than q, its parent node is used as the neighborhood of the test instance.

Next, we present a detailed description of our algorithm for learning such a decision tree, where D denotes a training dataset and q is a predetermined parameter for controlling the size of the tree. In the current version, we use 30 as the default value of q. This is the same as the leaf size of an NBTree [11].

Algorithm : Decision Tree (D, q)

```
Input : a training dataset $D$, a predetermined parameter $q$
Output : the built decision tree $T$
1. Create a Root node for $T$
2. If the number of instances in Root is less than $q$, return the
   single-node tree $T$
3. Otherwise Begin
      3.1 For each attribute $A_i$, use Eq. 9 to calculate
          $d_N(P_C, P_{A_i})$
      3.2 Let $A$ be the test attribute with the minimum distance $d_N$
      3.3 If $d_N(P_C, P_A) = 1$, create a leaf node
      3.4 Otherwise Begin
             i. For each possible value $v_i$ of $A$, create a child node,
                corresponding to the test $A = v_i$
             ii. For each child node, recursively call the algorithm
                 till all the nodes are leaf nodes
      3.5 End
4. End
5. Return the built decision tree $T$
```

When called to predict a test instance y, LVDM first sorts y down the built decision tree T from the root to a leaf node, and then uses only the instances falling into this leaf node to estimate the conditional probabilities in VDM. If the number of instances in this

Table 2
RRSE (%) comparisons for VDM versus OM and LVDM.

Dataset	VDM	OM	LVDM
anneal	70.84 ± 1.31	90.27 ± 1.12○	67.48 ± 1.78●
anneal.ORIG	72.86 ± 2.05	80.52 ± 1.52○	73.25 ± 2.33
audiology	99.28 ± 0.55	100.89 ± 0.38○	98.55 ± 0.79●
autos	90.27 ± 1.88	93.60 ± 1.49○	88.75 ± 2.31●
balance-scale	72.37 ± 3.13	72.29 ± 3.02	72.37 ± 3.13
breast-cancer	96.77 ± 5.26	94.98 ± 2.63	97.16 ± 4.83
breast-w	39.80 ± 4.02	37.11 ± 5.01●	35.44 ± 5.99●
car-evaluation	51.95 ± 1.99	78.61 ± 0.73○	58.98 ± 1.65○
cardiotocography-2	67.67 ± 2.02	86.02 ± 1.58○	64.47 ± 2.53●
colic	77.04 ± 4.60	89.17 ± 1.77○	74.41 ± 5.01●
colic.ORIG	84.13 ± 5.60	97.01 ± 0.95○	85.96 ± 5.26
connectionist-vowel	83.07 ± 0.77	84.36 ± 0.64○	81.96 ± 0.84●
credit-a	65.60 ± 2.98	72.79 ± 2.45○	65.57 ± 3.62
credit-g	92.08 ± 2.16	95.33 ± 1.12○	93.25 ± 3.01
diabetes	83.48 ± 4.18	83.61 ± 2.98	82.97 ± 4.17
glass	79.19 ± 3.74	78.72 ± 3.48	79.22 ± 3.14
heart-c	85.39 ± 4.57	88.30 ± 3.23○	83.58 ± 5.00●
heart-h	78.16 ± 7.64	76.77 ± 6.87	77.43 ± 7.04
heart-statlog	71.60 ± 4.81	70.43 ± 3.53	71.65 ± 6.01
hepatitis	80.37 ± 8.21	87.08 ± 5.33○	80.52 ± 10.06
ionosphere	75.74 ± 1.91	81.63 ± 1.57○	61.33 ± 2.86●
iris	36.06 ± 9.17	34.76 ± 8.91	36.51 ± 10.34
kr-vs-kp	36.40 ± 1.89	60.69 ± 0.94○	28.73 ± 1.57●
labor	62.07 ± 8.54	67.16 ± 6.45○	60.89 ± 10.47
letter	89.74 ± 0.12	92.39 ± 0.10○	88.19 ± 0.14●
libras-movement	97.11 ± 0.34	97.17 ± 0.33○	96.27 ± 0.38●
lymphography	89.11 ± 4.00	104.66 ± 1.67○	82.96 ± 5.04●
monks-problems-1	58.00 ± 15.41	67.64 ± 1.70	64.38 ± 13.04
monks-problems-2	83.94 ± 14.23	97.51 ± 2.14	87.55 ± 4.80
mushroom	17.69 ± 0.11	21.92 ± 0.23○	16.59 ± 0.18●
parkinsons	70.19 ± 6.50	74.41 ± 5.32○	62.76 ± 11.75●
primary-tumor	98.98 ± 0.40	99.47 ± 0.36○	98.64 ± 0.49●
segment	79.52 ± 0.46	83.55 ± 0.38○	68.48 ± 0.95●
sick	62.80 ± 3.83	70.18 ± 2.30○	62.11 ± 3.06
sonar	77.78 ± 4.15	79.94 ± 3.21○	71.35 ± 7.15●
soybean	91.41 ± 0.57	94.66 ± 0.40○	85.52 ± 0.75●
splice	93.54 ± 0.42	100.94 ± 0.33○	81.35 ± 0.75●
thyroid-disease	41.38 ± 2.02	51.59 ± 2.68○	43.41 ± 2.75
vehicle	77.37 ± 1.60	79.26 ± 1.42○	73.94 ± 1.78●
vote	50.15 ± 4.72	53.20 ± 4.52○	41.45 ± 6.82●
vowel	82.57 ± 0.65	87.51 ± 0.40○	82.47 ± 0.60
waveform-5000	85.56 ± 0.27	95.08 ± 0.09○	79.01 ± 0.52●
zoo	68.95 ± 3.22	70.96 ± 2.47○	68.58 ± 3.15
W/T/L	–	1/9/33	23/19/1

○, ● statistically significant upgradation or degradation.

leaf node is less than q, then the instances falling into its parent node are used to estimate the conditional probabilities in VDM. Consequently, the distance between each training instance x and the test instance y can be calculated. The detailed learning process of LVDM can be described as follows.

```
Algorithm : LVDM $(T, q, x, y)$
Input : the built decision tree $T$, the predetermined parameter
  $q$, a training instance $x$, and the test instance $y$
Output : the distance $d(x, y)$ between $x$ and $y$
1. Sort $y$ down the built decision tree $T$ from the root node to
   a leaf node
2. Let $l$ be the number of instances falling into this leaf node
    2.1 If $l < q$, then the parent node of this leaf node is
        marked as the neighborhood node $L$ of $y$
    2.2 Else, this leaf node is marked as the neighborhood
        node $L$ of $y$
3. Use the training instances falling into $L$ to estimate
   $P(c_j|a_i(x))$ and $P(c_j|a_i(y))$, where $i = 1, 2, \ldots, m$ and
   $j = 1, 2, \ldots, t$
4. Calculate the distance $d(x, y)$ between $x$ and $y$ using Eq. 3
5. Return the calculated distance $d(x, y)$
```

Compared to VDM, LVDM needs additional time to build a decision tree. The pseudo code of it is shown in the algorithm **Decision Tree** (D, q), which almost maintains the same order of computational overhead as the standard decision-tree learning algorithms such as ID3 [12] and C4.5 [13], requiring a time complexity of $O(Nm^2)$ [26,27], where N is the size of the training dataset D, and m is the number of attributes. However, the process of building such a decision tree occurs at the training stage. This is trivial compared to distance-based lazy learning algorithms. At the test stage, LVDM needs additional time to find the neighborhood node L of the test instance y from T, which only takes a time complexity of $O(m)$. However, this process can significantly reduce the time complexity of estimating the conditional probabilities $P(c_j|a_i(x))$ and $P(c_j|a_i(y))$. In a word, the proposed LVDM is almost as fast as VDM. The following experimental results in terms of running time of the chosen algorithm KNNDW can also support our conclusion.

4. Experiments and results

In order to validate the effectiveness of LVDM, we experimentally compare VDM with OM and the proposed LVDM in terms of the conditional log likelihood (CLL) and the mean root relative squared error (RRSE) of KNNDW [3]. We implemented KNNDW using VDM, OM, and LVDM respectively on a Waikato Environment for Knowledge Analysis (WEKA) platform [28]. In our implementation, the weighting function is $w_i = 1/(1 + d^2(x_i, y))$, where $d(x_i, y)$ is the distance between the test instance y and its ith neighbor x_i. Besides, we set the value of k to 10 and use Laplace correction to smooth the estimated probabilities.

We executed our experiments on 43 datasets downloaded from the University of California at Irvine repository [29], which includes all of the non-trivial datasets (They are denoted by boldface in Tables 1–5.) considered by Duch et al. [30]. In our experiment, missing attribute values were replaced with the modes of the nominal attribute values and the means of the numerical attribute values from the available data. Numerical attribute values were discretized using the Fayyad and Irani's MDL method [31] implemented in the WEKA platform [28]. Besides, we manually deleted three useless attributes in "colic.ORIG", "splice", and "zoo".

Table 3
Running time comparisons for VDM versus OM and LVDM.

Dataset	VDM	OM	LVDM
anneal	295.72 ± 16.32	29.44 ± 6.96●	318.00 ± 24.18
anneal.ORIG	294.44 ± 11.61	26.24 ± 7.45●	316.16 ± 18.39
audiology	126.80 ± 6.79	2.44 ± 5.71●	138.84 ± 10.30
autos	13.12 ± 5.86	2.52 ± 5.90●	15.08 ± 3.17
balance-scale	9.96 ± 7.63	8.12 ± 7.97	13.12 ± 7.37
breast-cancer	2.48 ± 5.80	1.92 ± 5.31	2.56 ± 5.99
breast-w	21.40 ± 7.44	4.44 ± 7.27●	25.00 ± 7.93
car-evaluation	135.56 ± 12.45	29.96 ± 4.52●	135.56 ± 13.28
cardiotocography-2	538.68 ± 21.70	108.08 ± 9.06●	555.72 ± 22.12
colic	14.00 ± 6.92	2.48 ± 5.80●	13.60 ± 5.15
colic.ORIG	16.84 ± 7.56	3.72 ± 6.76●	18.68 ± 6.30
connectionist-vowel	49.44 ± 5.94	4.36 ± 7.14●	48.00 ± 4.23
credit-a	31.88 ± 7.17	12.48 ± 6.38●	30.48 ± 3.03
credit-g	86.80 ± 11.17	21.88 ± 7.78●	84.92 ± 7.82
diabetes	20.56 ± 7.59	6.20 ± 7.76●	22.52 ± 8.02
glass	6.88 ± 7.93	0.60 ± 3.00	6.16 ± 7.71
heart-c	11.24 ± 7.17	1.28 ± 4.43●	12.56 ± 6.42
heart-h	11.36 ± 7.24	1.24 ± 4.29	11.32 ± 8.49
heart-statlog	5.00 ± 7.44	1.28 ± 4.43	6.36 ± 7.95
hepatitis	2.48 ± 5.80	1.28 ± 4.43	1.88 ± 5.20
ionosphere	17.52 ± 5.10	4.40 ± 7.21●	18.72 ± 6.28
iris	1.28 ± 4.43	0.00 ± 0.00	0.64 ± 3.20
kr-vs-kp	1469.36 ± 18.88	369.44 ± 13.37●	1565.76 ± 35.62○
labor	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
letter	256893.1 ± 5361	7352.2 ± 275.0●	255007.6 ± 2415.1
libras-movement	276.36 ± 12.47	10.72 ± 7.51●	276.44 ± 13.28
lymphography	3.76 ± 6.83	0.00 ± 0.00	6.24 ± 7.81
monks-problems-1	8.64 ± 7.83	2.52 ± 5.90	12.44 ± 6.36
monks-problems-2	11.40 ± 7.26	3.72 ± 6.76	10.60 ± 7.43
mushroom	6398.96 ± 77.80	1511.4 ± 19.96●	6429.68 ± 131.42
parkinsons	1.20 ± 4.15	0.64 ± 3.20	5.00 ± 7.44
primary-tumor	69.96 ± 10.31	1.88 ± 5.20●	71.08 ± 10.18
segment	1165.68 ± 20.14	95.72 ± 7.01●	1163.20 ± 26.71
sick	1684.00 ± 34.16	350.68 ± 12.78●	1730.20 ± 35.73
sonar	13.20 ± 8.72	3.08 ± 6.29	13.72 ± 5.19
soybean	473.76 ± 16.84	15.64 ± 0.49●	479.36 ± 19.55
splice	3309.52 ± 61.05	636.44 ± 19.97●	3253.92 ± 63.99
thyroid-disease	6692.00 ± 69.17	957.60 ± 27.50●	7919.88 ± 104.79○
vehicle	89.44 ± 11.60	13.68 ± 5.18●	94.56 ± 11.48
vote	12.96 ± 5.79	4.40 ± 7.21	15.68 ± 0.48
vowel	222.52 ± 12.20	16.24 ± 3.11●	232.48 ± 19.79
waveform-5000	5600.96 ± 73.76	835.72 ± 16.20●	5723.44 ± 405.20
zoo	1.80 ± 4.97	0.60 ± 3.00	3.20 ± 6.53
W/T/L	–	28/15/0	0/41/2

○, ● Statistically significant upgradation or degradation.

Tables 1 and 2 show the CLL and RRSE (%) comparisons, respectively, for KNNDW with VDM versus OM and LVDM. The experimental results on each dataset were obtained via 5 times averaging the 5-fold cross-validation results. The symbols ○ and ● in the tables denote statistically significant upgradation or degradation over VDM with a corrected paired two-tailed t-test with the $p = 0.05$ significance level [32], respectively. The *Win*/*Tie*/*Lose* (*W*/*T*/*L*) values are summarized at the bottom of the tables. Each entry's *W*/*T*/*L* in the tables implies that, compared to VDM, OM and LVDM win on *W* datasets, tie on *T* datasets, and lose on *L* datasets. Note that the meaning of the t-test results in Table 2 are opposite to those in Table 1. For the RRSE, a small number is better than a large number. CLL is the opposite. Thus, in terms of RRSE, ● represents statistically better than VDM and ○ represents worse. For additional insight into the results, we graphically show the averaged CLL and RRSE (%) comparisons in Figs. 1 and 2 respectively.

From these compared results, we can see that VDM is decidedly better than OM and LVDM significantly outperforms VDM in terms of CLL and RRSE. The highlights are summarized as follows.

Table 4
CLL comparisons for LVDM with different q values.

Dataset	LVDM-30	LVDM-10	LVDM-50
anneal	−87.92 ± 3.17	−88.09 ± 3.04	−87.53 ± 3.39
anneal.ORIG	−94.79 ± 4.35	−95.41 ± 3.52	−94.96 ± 4.32
audiology	−109.35 ± 3.53	−102.09 ± 3.74○	−110.17 ± 3.09●
autos	−54.18 ± 2.62	−51.91 ± 2.87○	−55.18 ± 2.32●
balance-scale	−67.67 ± 5.01	−67.64 ± 4.63	−67.67 ± 5.01
breast-cancer	−33.40 ± 2.83	−32.77 ± 2.72	−33.74 ± 2.75
breast-w	−17.92 ± 3.34	−18.91 ± 3.55	−18.18 ± 3.34
car-evaluation	−132.64 ± 4.22	−132.84 ± 4.57	−128.93 ± 4.26
cardiotocography-2	−147.58 ± 7.41	−150.32 ± 8.14	−147.15 ± 7.63
colic	−30.90 ± 3.33	−31.06 ± 3.63	−30.51 ± 3.20
colic.ORIG	−36.74 ± 3.58	−36.44 ± 3.59	−36.09 ± 3.80
connectionist-vowel	−145.93 ± 3.72	−140.07 ± 5.06○	−148.10 ± 3.58●
credit-a	−49.50 ± 4.09	−50.05 ± 4.51	−48.57 ± 3.65
credit-g	−109.06 ± 5.87	−108.51 ± 4.82	−107.68 ± 5.52
diabetes	−73.65 ± 6.19	−73.33 ± 5.78	−73.45 ± 5.91
glass	−41.93 ± 3.03	−41.31 ± 3.53	−42.23 ± 3.23
heart-c	−44.31 ± 4.12	−45.35 ± 3.96●	−43.91 ± 4.19
heart-h	−34.06 ± 5.03	−33.91 ± 4.89	−33.81 ± 5.31
heart-statlog	−22.19 ± 2.91	−22.69 ± 3.17	−22.42 ± 2.57
hepatitis	−11.02 ± 2.16	−10.65 ± 2.09	−10.79 ± 2.19
ionosphere	−21.84 ± 1.26	−20.83 ± 1.38○	−22.55 ± 1.53
iris	−6.89 ± 2.03	−6.93 ± 2.02	−6.87 ± 1.77
kr-vs-kp	−79.45 ± 3.49	−71.17 ± 3.63○	−81.92 ± 3.90
labor	−3.47 ± 0.81	−3.59 ± 1.05	−3.59 ± 0.67
letter	−8169.60 ± 37.31	−7616.30 ± 40.26○	−8350.61 ± 37.85●
libras-movement	−169.77 ± 2.62	−164.89 ± 3.18○	−171.50 ± 2.52●
lymphography	−21.60 ± 2.08	−21.24 ± 2.21	−22.15 ± 2.03
monks-problems-1	−38.89 ± 10.23	−26.06 ± 6.70○	−51.44 ± 5.02●
monks-problems-2	−61.20 ± 5.22	−68.55 ± 4.79●	−58.28 ± 9.05
mushroom	−139.88 ± 1.05	−139.33 ± 0.93○	−140.61 ± 1.03●
parkinsons	−10.74 ± 2.43	−10.27 ± 2.39	−11.09 ± 2.08
primary-tumor	−171.48 ± 3.25	−166.15 ± 3.52○	−172.47 ± 3.22
segment	−421.59 ± 10.48	−409.56 ± 8.21○	−429.35 ± 10.47●
sick	−87.13 ± 4.91	−89.76 ± 6.12	−86.73 ± 4.83
sonar	−17.06 ± 2.79	−15.79 ± 3.14○	−17.27 ± 2.28
soybean	−231.84 ± 5.06	−200.69 ± 6.08○	−245.25 ± 4.79●
splice	−469.20 ± 6.73	−500.30 ± 10.63●	−459.63 ± 6.98○
thyroid-disease	−118.61 ± 7.36	−112.76 ± 7.85○	−119.02 ± 7.04
vehicle	−131.59 ± 5.45	−130.66 ± 4.88	−131.45 ± 4.86
vote	−15.80 ± 3.10	−15.56 ± 3.18	−15.76 ± 3.07
vowel	−279.73 ± 4.93	−272.27 ± 6.05○	−285.24 ± 4.51●
waveform-5000	−739.38 ± 7.37	−790.47 ± 5.33●	−721.78 ± 7.91○
zoo	−16.71 ± 1.39	−15.19 ± 1.21○	−16.86 ± 1.42
Average	−296.94	−283.76	−301.45
W/T/L	–	16/23/4	2/31/10

○, ● Statistically significant upgradation or degradation.

1. In terms of CLL, VDM is notably better than OM with 33 wins and two losses. Further, the average CLL value (−316.19) of VDM is significantly higher than that of OM (−357.46).
2. In terms of CLL, LVDM significantly outperformed VDM with 23 wins and one loss. Additionally, the average CLL value (−296.94) of LVDM is remarkably higher than that of VDM.
3. In terms of RRSE, VDM is markedly better than OM with 33 wins and one loss. In addition, the average RRSE value (73.67%) of VDM is substantially lower than that of OM (79.63%).
4. In terms of RRSE, LVDM significantly outperformed VDM with 23 wins and one loss. Further, the average RRSE value (71.52%) of LVDM is significantly lower than that of VDM.
5. All of the above experimental results validate the effectiveness of the proposed LVDM. LVDM makes the attribute-independence assumption required by VDM hold as much as possible, and thus improves the class-probability estimation performance of distance-based learning algorithms.

Table 5
RRSE (%) comparisons for LVDM with different q values.

Dataset	LVDM-30	LVDM-10	LVDM-50
anneal	67.48 ± 1.78	67.58 ± 1.70	67.22 ± 1.94
anneal.ORIG	73.25 ± 2.33	73.98 ± 1.90	73.20 ± 2.30
audiology	98.55 ± 0.79	96.60 ± 0.94●	98.87 ± 0.64○
autos	88.75 ± 2.31	86.96 ± 2.58●	89.77 ± 2.03○
balance-scale	72.37 ± 3.13	72.34 ± 2.94	72.37 ± 3.13
breast-cancer	97.16 ± 4.83	95.92 ± 4.56	97.91 ± 4.44
breast-w	35.44 ± 5.99	36.92 ± 5.80	35.95 ± 6.01
car-evaluation	58.98 ± 1.65	58.85 ± 1.68	57.59 ± 1.66
cardiotocography-2	64.47 ± 2.53	65.05 ± 2.70	64.56 ± 2.57
colic	74.41 ± 5.01	74.48 ± 5.65	73.88 ± 4.85
colic.ORIG	85.96 ± 5.26	85.51 ± 5.28	84.89 ± 5.62
connectionist-vowel	81.96 ± 0.84	80.30 ± 1.31●	82.47 ± 0.78○
credit-a	65.57 ± 3.62	66.00 ± 3.95	64.91 ± 3.30
credit-g	93.25 ± 3.01	93.00 ± 2.54	92.62 ± 2.83
diabetes	82.97 ± 4.17	82.93 ± 3.85	82.88 ± 4.00
glass	79.22 ± 3.14	78.61 ± 3.67	79.69 ± 3.37
heart-c	83.58 ± 5.00	84.66 ± 4.64○	83.32 ± 5.29
heart-h	77.43 ± 7.04	77.18 ± 6.77	77.48 ± 7.53
heart-statlog	71.65 ± 6.01	72.42 ± 6.26	71.88 ± 5.35
hepatitis	80.52 ± 10.06	78.63 ± 10.00	79.35 ± 10.11
ionosphere	61.33 ± 2.86	59.30 ± 3.28	62.94 ± 2.99
iris	36.51 ± 10.34	36.61 ± 10.27	36.06 ± 9.17
kr-vs-kp	28.73 ± 1.57	24.97 ± 1.79●	29.48 ± 1.69
labor	60.89 ± 10.47	62.78 ± 12.70	62.07 ± 8.54
letter	88.19 ± 0.14	86.13 ± 0.16●	88.82 ± 0.14○
libras-movement	96.27 ± 0.38	95.35 ± 0.52●	96.63 ± 0.36○
lymphography	82.96 ± 5.04	81.49 ± 5.24	84.28 ± 4.93
monks-problems-1	64.38 ± 13.04	48.09 ± 10.32●	78.48 ± 4.84○
monks-problems-2	87.55 ± 4.80	94.20 ± 3.56○	84.41 ± 8.82
mushroom	16.59 ± 0.18	16.50 ± 0.13	16.69 ± 0.18○
parkinsons	62.76 ± 11.75	61.66 ± 12.13	64.41 ± 9.49
primary-tumor	98.64 ± 0.49	97.59 ± 0.58●	98.85 ± 0.46○
segment	68.48 ± 0.95	67.37 ± 0.76●	69.22 ± 0.92○
sick	62.11 ± 3.06	63.65 ± 3.72	61.91 ± 3.02
sonar	71.35 ± 7.15	67.77 ± 8.18	71.94 ± 6.25
soybean	85.52 ± 0.75	80.51 ± 1.00●	87.60 ± 0.63○
splice	81.35 ± 0.75	84.83 ± 1.14○	80.25 ± 0.78●
thyroid-disease	43.41 ± 2.75	40.38 ± 3.16●	43.62 ± 2.57
vehicle	73.94 ± 1.78	73.40 ± 1.60	73.91 ± 1.63
vote	41.45 ± 6.82	41.11 ± 6.60	41.36 ± 6.79
vowel	82.47 ± 0.60	81.27 ± 0.78●	83.37 ± 0.52○
waveform-5000	79.01 ± 0.52	82.29 ± 0.35○	77.93 ± 0.57●
zoo	68.58 ± 3.15	64.61 ± 2.71●	68.95 ± 3.22
Average	71.52	70.69	71.95
W/T/L	–	13/26/4	2/30/11

○, ● Statistically significant upgradation or degradation.

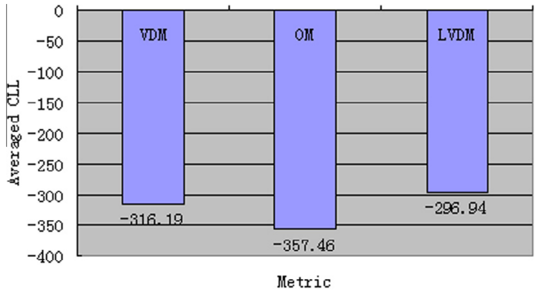

Fig. 1. Averaged CLL comparisons for VDM versus OM and LVDM.

We also compare VDM with OM and LVDM in terms of running time. The running time is measured by the averaged CPU time in millisecond. Our experiments are performed on a notebook PC with Microsoft Windows 8 Pro 64 bit with Intel (R) Core (TM) i7-3630QM Quad-Core CPU (2.4 GHz) and 8 GB memory. The detailed results are shown in Table 3 and Fig. 3. From the experimental

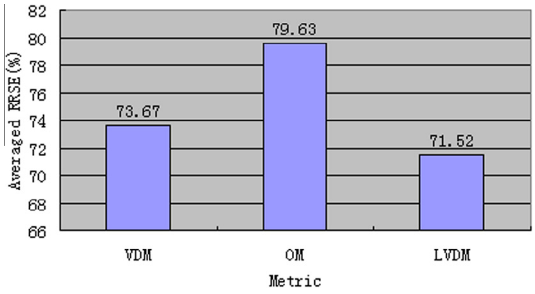

Fig. 2. Averaged RRSE (%) comparisons for VDM versus OM and LVDM.

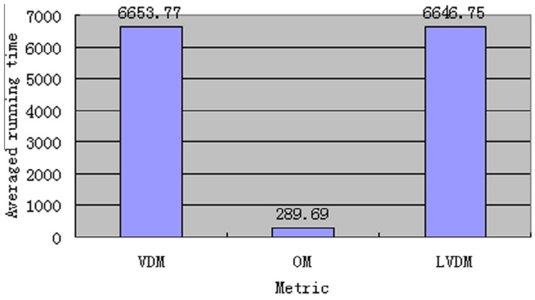

Fig. 3. Averaged running time comparisons for VDM versus OM and LVDM.

results, we can see that LVDM maintains almost the same order of computational overhead as VDM. Please note that the meaning of the *t*-test results in Table 3 are also opposite to those in Table 1.

In the current version, we use 30 as the default value of the predetermined parameter q, because this is the same as the leaf size of an NBTree [11]. To observe the influence of the predetermined parameter q on the resulting algorithm, we designed another group of experiments to compare LVDM with different q values. The compared results in Tables 4 and 5 show that LVDM is relatively insensitive to the size of the predetermined parameter q, provided that it is not extremely small (<10) or large (>50).

Besides, we repeated our experiments using the unsupervised ten-bin discretization method and got almost the same conclusions. Due to the total number of pages, figures and tables limit, we did not present the detailed experimental results here.

5. Conclusions

Several studies show that local learning for NB can ensure the attribute-independence assumption to likely be true and consequently improve the performance of NB. Similar to NB, VDM also assumes that all attributes are independent given the class. Therefore, in this paper, we proposed a local learning approach for VDM and called it local value difference metric (LVDM). The experimental results on a large suite of benchmark datasets showed that LVDM significantly outperformed VDM in terms of CLL and RRSE.

VDM has a close relationship with NB, and the definition of VDM follows naturally from NB. Therefore, those techniques proposed to improve NB can be considered to improve VDM. This paper showed that the local learning technique that was effective for NB was also effective for VDM. Our future research will involve using other learning techniques designed to improve NB, to improve VDM.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (61203287), the Program for New Century Excellent Talents in University (NCET-12-0953), and the Fundamental Research Funds for the Central Universities (CUG130504, CUG130414).

References

[1] D. Aha, D. Kibler, M.K. Albert, Instance-based learning algorithms, Mach. Learn. 6 (1991) 37–66.
[2] D. Aha, Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms, Int. J. Man-Mach. Stud. 36 (2) (1992) 267–287.
[3] T.M. Mitchell, Machine Learning, McGraw-Hill, 1997.
[4] C.G. Atkeson, A.W. Moore, S. Schaal, Locally weighted learning, Artif. Intell. Rev. 11 (1–5) (1997) 11–73.
[5] E. Frank, M. Hall, B. Pfahringer, Locally weighted naive Bayes, in: Proceedings of the Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, 2002, pp. 249–256.
[6] L. Jiang, Z. Cai, H. Zhang, D. Wang, Naive Bayes text classifiers: a locally weighted learning approach, J. Exp. Theor. Artif. Intell. 25 (2) (2013) 273–286.
[7] D.R. Wilson, T.R. Martinez, Improved heterogeneous distance functions, J. Artif. Intell. Res. 6 (1) (1997) 1–34.
[8] C. Stanfill, D. Waltz, Toward memory-based reasoning, Commun. ACM 29 (1986) 1213–1228.
[9] S. Kasif, S. Salzberg, D. Waltz, J. Rachlin, D. Aha, A probabilistic framework for memory-based reasoning, Artif. Intell. 104 (1998) 287–311.
[10] X. Wu, V. Kumar, J.R. Quinlan, et al., Top 10 algorithms in data mining, Knowl. Inform. Syst. 14 (1) (2008) 1–37.
[11] R. Kohavi, Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1996, pp. 202–207.
[12] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[13] J.R. Quinlan, C4.5 (1993): Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[14] R.L.D. Mantaras, A distance-based attribute selection measure for decision tree induction, Mach. Learn. 6 (1991) 81–92.
[15] S. Cost, S. Salzberg, A weighted nearest neighbor algorithm for learning with symbolic features, Mach. Learn. 10 (1) (1993) 57–78.
[16] C. Li, H. Li, One dependence value difference metric, Knowl.-Based Syst. 24 (5) (2011) 589–594.
[17] L. Jiang, C. Li, An augmented value difference measure, Pattern Recognit. Lett. 34 (10) (2013) 1169–1174.
[18] M. Saar-Tsechansky, F. Provost, Active sampling for class probability estimation and ranking, Mach. Learn. 54 (2) (2004) 153–178.
[19] L. Jiang, H. Zhang, Learning naive Bayes for probability estimation by feature selection, in: Proceedings of the 19th Canadian Conference on Artificial Intelligence, CAI 2006, LNAI, vol. 4013, Springer Press, 2006, pp. 503–514.
[20] J. Hu, N. Zhong, Y. Shi, Developing mining-grid centric e-finance portals for risk management and decision making, Int. J. Pattern Recognit. Artif. Intell. 21 (4) (2007) 639–658.
[21] D. Grossman, P. Domingos, Learning Bayesian network classifiers by maximizing conditional likelihood, in: Proceedings of the 21st International Conference on Machine Learning, ICML 2004, ACM Press, 2004, pp. 361–368.
[22] Y. Guo, R. Greiner, Discriminative model selection for belief net structures, in: Proceedings of the 20th National Conference on Artificial Intelligence, AAAI 2005, AAAI Press, 2005, pp. 770–776.
[23] M. Hall, A decision tree-based attribute weighting filter for naive Bayes, Knowl.-Based Syst. 20 (2007) 120–126.
[24] L. Jiang, C. Li, An empirical study on attribute selection measures in decision tree learning, J. Comput. Inform. Syst. 6 (1) (2010) 105–112.
[25] Y.Y. Yao, Information-theoretic measures for knowledge discovery and data mining, in: Karmeshu (Ed.), Entropy Measures, Maximum Entropy and Emerging Applications, Springer, Berlin, 2003, pp. 115–136.
[26] J.K. Martin, D.S. Hirschberg, On the complexity of learning decision trees, in: Proceedings of the Fourth Int. Symp. on Artif. Intell. and Math., Fort Lauderdale, FL, 1996, pp. 112–115.
[27] J. Su, H. Zhang, A fast decision tree learning algorithm, in: Proceedings of the 21th National Conference on Artificial Intelligence (AAAI'06), Boston, MA, USA, 2006, pp. 500–505.
[28] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, third ed., Morgan Kaufmann, 2011, ISBN 978-0-12-374856-0.
[29] A. Frank, A. Asuncion, UCI machine learning repository, University of California, School of Information and Computer Science, Irvine, CA, 2010 (<http://archive.ics.uci.edu/ml>).
[30] W. Duch, N. Jankowski, T. Maszczyk, Make it cheap: learning with O(nd) complexity, in: IEEE World Congress on Computational Intelligence, IEEE Press, 2012, pp. 2520–2527.
[31] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1993, pp. 1022–1027.
[32] C. Nadeau, Y. Bengio, Inference for the generalization error, Mach. Learn. 52 (3) (2003) 239–281.