A Survey of Distance Metrics for Nominal Attributes

Chaoqun Li and Hongwei Li
Department of Mathematics, China University of Geosciences, Wuhan, Hubei, China 430074
Email: {chqli, hwli}@cug.edu.cn

Abstract— **Many distance-related algorithms, such as k-nearest neighbor learning algorithms, locally weighted learning algorithms etc, depend upon a good distance metric to be successful. In this kind of algorithms, a key problem is how to measure the distance between each pair of instances. In this paper, we provide a survey on distance metrics for nominal attributes, including some basic distance metrics and their improvements based on attribute weighting and attribute selection. The experimental results on the whole 36 UCI datasets published on the main web site of Weka platform validate their effectiveness.**

Index Terms— **distance metric, attribute weighting, attribute selection, nominal attributes, classification**

I. Introduction

Although many distance metrics reviewed in this paper can be used to address the regression and clustering problems directly, we focus our attention on the classification problem. In classification, an arbitrary instance x is often represented by an attribute vector $< a_1(x), a_2(x), \ldots, a_n(x) >$, where $a_i(x)$ denotes the value of the ith attribute A_i of x.

The distance measure problem includes two aspects: 1) How to choose an appropriate distance metric according to different data characteristics. Some are appropriate for numeric attributes while others are appropriate for nominal attributes. In this paper, we focus our attention on the distance metrics for nominal attributes. 2) How to overcome the *curse of dimensionality* problem [1], [2]. The distance-related algorithms are especially sensitive to this problem when there are a lot of redundant and/or irrelevant attributes in the data. An effective approach to overcome the *curse of dimensionality* problem is to weight each attribute differently when measuring the distance between each pair of instances. This approach is widely known as *attribute weighting*. Another drastic approach to overcome the *curse of dimensionality* is to completely eliminate the least relevant attributes from the attribute space when measuring the distance between each pair of instances. This approach is widely known as *attribute selection*.

The rest of the paper is organized as follows. In Section II, we provide a survey on some basic distance metrics for nominal attributes and their improvements based on attribute weighting and attribute selection. In Section III, we use the k-nearest neighbor algorithm (KNN) [3] and the whole 36 UCI datasets published on the main web site of Weka platform [4] to experimentally test their effectiveness. In Section IV, we draw conclusions.

Corresponding author: Hongwei Li (hwli@cug.edu.cn)

Chaoqun Li is supported by the National Natural Science Foundation of China under grant no. 60905033, the Research Foundation for Outstanding Young Teachers, China University of Geosciences (Wuhan) under grant no. CUGQNL0830, and the Fundamental Research Funds for the Central Universities under grant no. CUGL090248.

II. Distance Metrics

A. Basic Distance Metrics

When all attributes are nominal, the simplest distance metric is the Overlap Metric. We simply denote it OM in this paper, which can be defined as:

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

where n is the number of attributes, $a_i(x)$ and $a_i(y)$ are the values of the ith attribute A_i of the instances x and y respectively, $\delta(a_i(x), a_i(y))$ is 0 if $a_i(x) = a_i(y)$ and 1 otherwise.

OM is widely used by instance-based learning [5], [6] and locally weighted learning [7], [8]. Obviously, it is a little rough to measure the distance between each pair of instances, because it fails to make use of additional information provided by nominal attribute values that can aid in generalization [9].

In order to find reasonable distance between each pair of instances with nominal attribute values only, the Value Difference Metric (VDM) was introduced by Standfill and Waltz [10]. A simplified version of it, without the weighting schemes, can be defined as:

$$d(x, y) = \sum_{i=1}^{n} \sum_{c=1}^{C} |P(c|a_i(x)) - P(c|a_i(y))| \quad (2)$$

where C is the number of classes, $P(c|a_i(x))$ is the conditional probability that the class of x is c given that the attribute A_i has the value $a_i(x)$, $P(c|a_i(y))$ is the conditional probability that the class of y is c given that the attribute A_i has the value $a_i(y)$.

VDM assumes that two values of an attribute are more closer if they have more similar classifications. Our experimental results in Section III show that VDM is much more accurate than OM.

Another probability-based metric is SFM, which was originally presented by Short and Fukunaga [11], [12] and

doi:10.4304/jsw.5.11.1262-1269

then was extended by Myles and Hand [13]. It is defined by Equation 3.

$$d(x,y)=\sum_{c=1}^{C}|P(c|x)-P(c|y)| \tag{3}$$

where the class membership probabilities $P(c|x)$ and $P(c|y)$ can only be estimated by naive Bayes [14], [15] in many realistic data mining applications, because more complicated probability estimators will lead to more higher time complexity.

After this, Minimum Risk Metric (simply MRM, defined by Equation 4) was presented by Blanzieri and Ricci [16]. Different from SFM that minimizes the expectation of difference between the finite error and the asymptotic error, MRM directly minimizes the risk of misclassification.

$$d(x,y)=\sum_{c=1}^{C}P(c|x)(1-P(c|y)) \tag{4}$$

Besides, Cleary and Trigg [17] presented a distance metric based on entropy. The approach computing the distance between two instances is motivated by information theory, and the distance between instances is defined as the complexity of transforming one instance into another. The advantage of the measure is to provide a consistent approach to handling of symbolic attributes, real valued attributes and missing values.

Daniel Tunkelang and Daniel Tunkelang Endeca [18] presented a data-driven difference measure for categorical data for which the difference between two data points is based on the frequency of the categories or combinations of categories that they have in common.

According to our experiments in Section III, no any one distance metric can perform better than the other on all application domains because different distance metrics have different biases. Therefore, we can choose different distance metrics for different data mining applications.

B. Improving Basic Distance Metrics via Attribute Weighting

In above basic distance metrics, all attributes are considered to have identical contributions to the distance metrics. However, this assumption is unrealistic in many data mining application domains. To relax this assumption, one way is to assign different weights to different attributes when measuring the distance between each pair of instances. This is the well-known *attribute weighting*. In this Section, we take the Overlap Metric for example to discuss different kinds of attribute weighting methods.

First of all, let's give the definition of the attribute weighted Overlap Metric as follows.

$$d(x,y)=\sum_{i=1}^{n}w_i\delta(a_i(x),a_i(y)) \tag{5}$$

where w_i is the weight of the ith attribute A_i.

Now, the only left thing is how to define the weight of each attribute.

When all attributes are nominal, many times, the mutual information is used to define the correlation between each attribute variable and the class variable. In fact, it has already been widely used in many papers [19]–[21]. We call the resulting OM mutual information weighted overlap metric, simply MIWOM. Now, let's define it as follows.

$$w_i=\sum_{a_i,c}P(a_i,c)\log\frac{P(a_i,c)}{P(a_i)P(c)} \tag{6}$$

Instead of mutual information, Huang [22] used the frequencies with which the attribute values $a_i(x)$ and $a_i(y)$ appear in the training data to weight the attribute A_i. It can be defined as follows.

$$w_i=\frac{F(a_i(x))+F(a_i(y))}{F(a_i(x))F(a_i(y))} \tag{7}$$

where $F(a_i(x))$ and $F(a_i(y))$ are the frequencies with which the attribute values $a_i(x)$ and $a_i(y)$ appear in the training data respectively. We call the resulting OM frequency weighted overlap metric, simply FWOM.

In 2007, Hall [23] proposed to obtain the weights of all of the attributes by building decision trees. For detail, this algorithm constructs an ensemble of unpruned decision trees at first, and then the minimum depth that an attribute is tested at the built decision trees is used to weight the attribute. The detailed Equation is described as follows.

$$w_i=\frac{\sum_{j=1}^{s}1/\sqrt{d_j}}{s} \tag{8}$$

where d_j is the minimum depth that the attribute A_i is tested at the jth decision tree, and s is the number of trees that the attribute A_i appears in.

To classification problem, there are more elaborate attribute weighting methods. In these methods, the weights are class sensitive in that an attribute may be more important to one class than to another. To cater for this, Aha [6] described IB4 which maintains a separate description and a set of attribute weights for each class, and the weights are adjusted using a simple performance feedback algorithm to reflect the relative relevances of the attributes. All weights are updated after each training instance x is classified. The updating is on the basis of the most similar instance y of x, the difference of the ith attribute value between x and y, and whether the classification is indeed correct. The weights increase when they correctly predict classifications and decrease otherwise.

Wettschereck and Aha [24] investigated those methods that automatically assign weights for all of attributes using little or no domain-specific knowlege, and introduced a five-dimensional framework to categorize automatic weight-setting methods.

Instead of considering an continuous weight space, Kohavi etc [25] described a method searching a discrete weight space. They thought that a large space of weight will lead to increase variance and over-fitting. Their experimental results show that, to many datasets, restricting the number of possible weights to two (0 and 1) is superior.

In fact, this is the so-called attribute selection. We will review it in the next subsection.

Although attribute selection can be viewed as a special case of attribute weighting with the zero and one weights only, there are two main difference between attribute weighting and attribute selection at least: 1) The main problem focused on is different. Most attribute weighting methods focus on how to calculate the relevance degree between the attribute variable and the class variable. In contrast, most attribute selection methods focus on how to search and evaluate attribute subsets. 2) The main motivation is different. Most attribute weighting methods focus on how to improve the accuracy of the resulted algorithms. In contrast, most attribute selection methods focus on how to reduce the test time of the resulted algorithms and the dimensionality of the available data. According to our experiments in Section III, the attribute weighting methods can enhance the performance of the resulted algorithms indeed while the attribute selection methods can significantly reduce the test time of the resulted algorithms and the size of dimensionality of the available data.

C. Improving Basic Distance Metrics via Attribute Selection

The extreme of attribute weighting is attribute selection, which assumed that some attributes are completely irrelevant and redundant to class variable. In fact, in many applications, there exist such attributes indeed [26]–[28]. Therefore, these attributes should be removed from attribute space. In the same way, we also take the Overlap Metric for example to discuss different kinds of attribute selection methods.

First of all, let's give the definition of the attribute selected Overlap Metric as follows.

$$d(x,y) = \sum_{i=1}^{k} \delta(a_i(x), a_i(y)) \tag{9}$$

where $\{a_1, a_2, \ldots, a_k\}$ is the selected attribute subset.

Now, the only left thing is how to search and evaluate the attribute subset $\{a_1, a_2, \ldots, a_k\}$.

The attribute selection methods can be broadly divided into two main categories: wrapper methods and filter methods. 1) wrapper methods use a learning algorithm to measure the merit of the selected attribute subsets. Thus, wrapper methods select different attribute subsets for different learning algorithms. Because wrapper methods consider how a learning algorithm and the training instances interact, wrapper models have more higher accuracy and computational cost than filter methods in many cases. 2) Filter methods have one common characteristic: the attribute selection process is independent of the learning algorithm. Therefore, filter methods can be viewed as a data preprocessing step. Generally, filter methods can be operated efficiently and effectively, and are adaptive for high dimensional data.

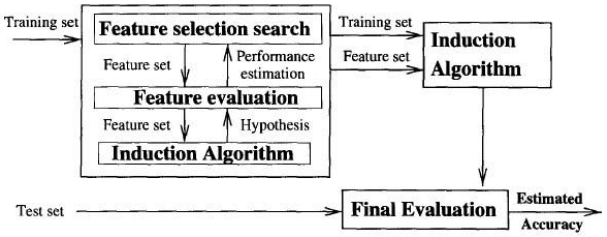

Figure 1. The basic framework of wrapper methods for attribute subset selection.

1) Wrapper Methods: In 1994, John and Kohavi etc [29] presented the define of filter methods and wrapper methods. Wrapper methods are those attribute selection methods utilizing the induction algorithm itself, their experiments show the efficiency of wrapper methods. To wrapper methods, the merit of an attribute subset is measured by the accuracy of the resulted learning algorithms, so the key problem is how to design search strategies. Research strategies mainly include exhaustive search and heuristic search. Exhaustive search can find the optimal attribute subset for a learning algorithm, but is not practical for high dimensional datasets. Heuristic search only searches a restricted space of attribute subsets and is more practical, but the selected attribute subset by heuristic search may be suboptimal. Heuristic search mainly includes greedy stepwise search, best first search, and genetic algorithms etc.

In 1997, Kohavi and John [30] in detail investigated the strengths and weaknesses of the wrapper approach, compared greedy hill climbing and best first research, and presented some improvements to reduce the running time. Simultaneously, they pointed that overuse the accuracy estimates in attribute subset selection may cause the over-fitting problem. To the over-fitting problem of the wrapper approach, Loughrey and Cunningham [31] outlined a set of experiments to prove the case, and presented a modified genetic algorithm to address the over-fitting problem by stopping the search before over-fitting occurs. Figure 1 shows the basic framework of wrapper methods for attribute subset selection, and the figure is described in [30] originally.

2) Filter Methods: To filter methods, except search strategy, another very important problem is how to evaluate the merit of the selected attribute subsets.

Almuallim and Dietterich [32] presented a method simply called FOCUS. FOCUS exhaustively searches the space of attribute subsets until it finds the minimum combination of attributes that is sufficient to determine the class label. That is referred to as the *min-features bias*. However, John etc [29] pointed out that this bias has some severe implications when it is applied blindly. For example, this bias favors the attributes with many values over those with few values. this is alike to the bias in the information gain measure in decision trees. Moreover this method is not practical for the high dimensional data because it executes an exhaustive search.

Kira and Rendell [33] presented a method called RELIEF, which uses the statistical method to remove irrelevant attributes. The method can be divided into four steps: The first step is randomly sampling m instances; the second step is finding the nearest instance y of the same class and the nearest instance z of opposite class to each sampled instance x; the third step is updating each attribute weight based m triplets of x, y, z; The last step is choosing a threshold τ, those attributes whose relevance are less than τ are considered to be irrelevant attributes and be removed from the attribute set. Similar to RELIEF, Scherf and Brauer [34] present another method (simply EUBAFES) by using an attribute weighting approach to attribute selection.

In the earlier research, researchers focused their attention on removing irrelevant attributes and ignored redundant attributes because Kira and Rendell [33] pointed out that Relief does not help with redundant attributes. The later researchers presented some more elaborate attribute selection methods to discriminate the irrelevant and redundant attributes.

M. A. Hall [35], [36] presented a method simply called CFS (Correlation based Feature Selection). The central hypothesis of CFS is that good attribute sets contain attributes that are highly correlated with the class variable, yet uncorrelated with each other. This method heuristically searches an attribute subset through a correlation based approach, and uses Equation 10 to measure the merit of an attribute subset *S* containing k attributes.

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{10}$$

where $\overline{r_{cf}}$ is the average attribute-class correlation, and $\overline{r_{ff}}$ is the attribute-attribute inter-correlation. Obviously, the heuristic merit tries to search an attribute subset with bigger $\overline{r_{cf}}$ by removing irrelevant attributes and smaller $\overline{r_{ff}}$ by removing redundant attributes.

Yu and Liu [37] presented another filter method using a correlation-based approach, in which an entropy-based measure is used to indicate the correlation between each pair of attributes. Besides, they discussed the definition of strong relevant attributes, weak relevant attributes, irrelevant attributes and redundant attributes in detail, and introduced a framework to discriminate the irrelevant and redundant attributes. The central assumption of the method is that an optimal attribute subset should contains all strong relevant attributes and weak relevant but non-redundant attributes.

Peng etc [38] introduced a criteria called minimal-redundancy-maximal-relevance (simply MRMR) to search a set of sequential attribute subsets candidates, then a wrapper method is used to find an optimal attribute subset from these attribute subsets candidates. The MRMR criteria is defined as:

$$max\Phi(D(S,c), R(S)), \Phi = D(S,c) - R(S) \tag{11}$$

where $D = I(\{a_i, i = 1, \ldots, s\}; c)$ is a feature set S with s features dependency on the target class c, $R = \frac{1}{|S|^2}\sum_{a_i,a_j \in S} I(a_i, a_j)$ is the dependency among these k features in the feature set S. The MRMR criteria is a balance between maximizing $D(S,c)$ and minimizing R(S).

Seen from above, almost all methods, except for RELIEF, involve search strategies. Therefore, these methods have relatively higher computational time. In fact, some researchers present some methods without involving the search process. For example, Cardie [39] applied decision tree algorithms to select attribute subsets for case-based learning, and only those attributes that appeared in the final decision trees are used in a k-nearest neighbor classifier.

Ratanamahatana and Gunopulos [40] also used decision trees to select attribute subsets for naive Bayes. It is known that the performance of NB suffers in domains that involve correlated attributes. C4.5 decision tree learning algorithm, on the other hand, typically perform better than NB on such domains. Therefore, in their method, only those attributes appeared in the decision trees built by C4.5 are selected for NB.

Mitra etc [41] presented another unsupervised attribute selection method without involving the search process. The method uses a cluster technique to partition the original attribute set into a number of homogeneous subsets and selects a representative attribute from each subset. Besides, Liu etc [42] presented a discretization-based method for attribute subset selection.

III. Experiments and Results

In this section, we use the 10-nearest neighbor algorithm to validate the effectiveness of some basic distance metrics and their improvements based on attribute weighting and attribute selection.

Experiments are performed on the whole 36 UCI datasets published on the main web site of Weka platform [4], which represent a wide range of domains and data characteristics listed in Table I. In our experiments, the following four data preprocessing steps are used:

1) Replacing missing attribute values: The unsupervised filter named *ReplaceMissingValues* in Weka is used to replace all missing attribute values in each data set.
2) Discretizing numeric attribute values: The unsupervised filter named *Discretize* in Weka is used to discretize all numeric attribute values in each data set.
3) Sampling large data sets: For saving the time of running experiments, the unsupervised filter named *Resample* with the size of 20% in Weka is used to randomly sample each large data set having more than 5000 instances. In these 36 data sets, there are three such data sets: "letter", "mushroom", and "waveform-5000".
4) Removing useless attributes: Apparently, if the number of values of an attribute is almost equal to the number of instances in a data set, the attribute is useless. Thus, we used the unsupervised filter

named *Remove* in Weka to remove this type of attributes. In these 36 data sets, there are only three such attributes: the attribute "Hospital Number" in the data set "colic.ORIG", the attribute "instance name" in the data set "splice" and the attribute "animal" in the data set "zoo".

Now, we introduce established algorithms and their abbreviations used in our implements and experiments.

1) OM: the Overlap Metric defined by Equation 1.
2) VDM: the Value Difference Metric defined by Equation 2.
3) SFM: the Short and Fukunaga Metric defined by Equation 3.
4) MIWOM: the Overlap Metric with the weights defined by Equation 5 and Equation 6.
5) FWOM: the Overlap Metric with the weights defined by Equation 5 and Equation 7.
6) Wrapper: the Overlap Metric with Wrapper-based attribute selection method [30]. Besides, the greedy stepwise search strategy is used.
7) Filter: the Overlap Metric with Filter-based attribute selection method [35], [36]. Besides, the greedy stepwise search strategy is used.

Three groups of experiments are designed: The first one is used to compare OM with VDM and SFM in terms of classification accuracy. The second one is used to validate the effectiveness of attribute weighting in terms of classification accuracy. The third one is used to validate the effectiveness of attribute selection in terms of classification accuracy, test time (the averaged CPU time in millisecond, the experiments are performed on a dual-processor 2.26 Ghz P8400 Windows notebook PC with 2.93Gb RAM.), and the size of dimensionality.

In all experiments, the classification accuracy, test time, and size of dimensionality on each dataset are obtained via 10-fold cross-validation. Finally, we conducted a two-tailed t-test with 95% confidence level [43] to compared each pair of algorithms.

Table II-VI respectively show the compared results. The symbols $\circ$ and $\bullet$ in the tables statistically significant upgradation or degradation over OM with a 95% confidence level. The averages and the $w/t/l$ values are summarized at the bottom of these tables, each entry $w/t/l$ means that the improved metrics win on w datasets, tie on t datasets, and lose on l datasets compared to OM. From the experimental results, we can see that:

1) No any one distance metric can perform better than the other on all datasets because different distance metrics have different biases. Therefore, we can choose different distance metrics for different data mining applications. Generally speaking, VDM and SFM are better than OM when high classification accuracy is the sole concern. when the computational cost and/or comprehensibility are also important, OM should be considered firstly.
2) The attribute weighting methods can enhance the performance of the resulted algorithms indeed while the attribute selection methods can significantly reduce the test time of the resulted algorithms and the size of dimensionality of the available data.

TABLE I.
DESCRIPTIONS OF UCI DATASETS USED IN THE EXPERIMENTS.

Dataset	Instances	Attributes	Classes
anneal	898	38	6
anneal.ORIG	898	38	6
audiology	226	69	24
autos	205	25	7
balance-scale	625	4	3
breast-cancer	286	9	2
breast-w	699	9	2
colic	368	22	2
colic.ORIG	368	27	2
credit-a	690	15	2
credit-g	1000	20	2
diabetes	768	8	2
Glass	214	9	7
heart-c	303	13	5
heart-h	294	13	5
heart-statlog	270	13	2
hepatitis	155	19	2
hypothyroid	3772	29	4
ionosphere	351	34	2
iris	150	4	3
kr-vs-kp	3196	36	2
labor	57	16	2
letter	20000	16	26
lymph	148	18	4
mushroom	8124	22	2
primary-tumor	339	17	21
segment	2310	19	7
sick	3772	29	2
sonar	208	60	2
soybean	683	35	19
splice	3190	61	3
vehicle	846	18	4
vote	435	16	2
vowel	990	13	11
waveform-5000	5000	40	3
zoo	101	17	7

IV. CONCLUSIONS

In this paper, we provide a survey on distance metrics for nominal attributes, including some basic distance metrics and their improvements based on attribute weighting and attribute selection. According to our experimental results on a large number of UCI datasets, we can draw conclusions: 1) Generally, VDM and SFM are better than OM when high classification accuracy is the sole concern. when the computational cost and/or comprehensibility are also important, OM should be considered firstly. 2) The attribute weighting methods, especially the attribute weighting method based on frequency, can really improve the classification accuracy of the resulted algorithms. 3) Although the attribute selection methods can't achieve significant improvements in terms of accuracy, it can significantly reduce the test time of the resulted algorithms and the size of dimensionality of the available data.

ACKNOWLEDGMENT

We thank anonymous reviewers for their very useful comments and suggestions.

TABLE II.
EXPERIMENTAL RESULTS FOR OM VERSUS VDM AND SFM: CLASSIFICATION ACCURACY.

Dataset	OM	VDM		SFM	
anneal	95.88	97.55	○	97.10	
anneal.ORIG	84.41	88.53	○	88.86	○
audiology	58.79	61.42		72.09	○
autos	62.52	64.93		68.71	
balance-scale	83.84	86.73		94.08	○
breast-cancer	73.09	74.11		70.30	
breast-w	93.99	96.28	○	96.85	○
colic	83.13	84.76		80.44	
colic.ORIG	69.82	76.63	○	74.76	
credit-a	86.09	84.64		83.62	
credit-g	71.90	74.40		73.30	
diabetes	69.02	75.14	○	74.24	
glass	57.06	61.26		60.80	
heart-c	81.09	83.77		81.81	
heart-h	82.02	82.37		85.08	
heart-statlog	82.22	82.59		83.70	
hepatitis	84.50	81.25		81.88	
hypothyroid	93.08	93.16		93.08	
ionosphere	89.74	90.03		90.89	
iris	93.33	95.33		94.67	
kr-vs-kp	95.06	96.21	○	87.27	●
labor	85.67	88.00		93.33	
letter	71.22	80.35	○	69.75	●
lymph	80.86	80.29		82.38	
mushroom	99.75	99.75		96.18	●
primary-tumor	42.47	43.93		46.59	
segment	89.65	93.33	○	89.78	
sick	97.03	97.67	○	96.55	
sonar	81.33	79.83		77.98	
soybean	89.01	93.41	○	92.23	○
splice	83.26	93.54	○	95.17	○
vehicle	68.68	71.16		59.58	●
vote	92.90	94.73		91.28	
vowel	67.68	82.42	○	70.00	
waveform-5000	74.60	82.80	○	82.90	○
zoo	89.18	89.18		91.18	
Average	80.66	83.37		82.46	
w/t/l	-	13/23/0		7/25/4	

○, ● statistically significant upgradation or degradation

TABLE III.
EXPERIMENTAL RESULTS FOR OM VERSUS THE IMPROVED OM BY ATTRIBUTE WEIGHTING: CLASSIFICATION ACCURACY.

Dataset	OM	MIWOM		FWOM	
anneal	95.88	96.99	○	97.66	○
anneal.ORIG	84.41	87.64	○	96.77	○
audiology	58.79	70.28	○	57.87	
autos	62.52	66.40		59.50	
balance-scale	83.84	78.24	●	98.88	○
breast-cancer	73.09	70.95		79.73	○
breast-w	93.99	95.56	○	76.53	●
colic	83.13	82.86		84.23	
colic.ORIG	69.82	74.45		71.22	
credit-a	86.09	85.94		86.52	
credit-g	71.90	75.50		79.30	○
diabetes	69.02	73.31		80.74	○
glass	57.06	58.48		68.66	○
heart-c	81.09	82.10		83.82	
heart-h	82.02	81.69		84.05	
heart-statlog	82.22	82.22		86.67	
hepatitis	84.50	81.29		87.71	
hypothyroid	93.08	93.21		98.25	○
ionosphere	89.74	89.46		85.76	
iris	93.33	92.67		94.67	
kr-vs-kp	95.06	95.65		97.50	○
labor	85.67	91.33		91.33	
letter	71.22	74.83	○	94.10	○
lymph	80.86	82.29		86.48	
mushroom	99.75	99.69		99.88	
primary-tumor	42.47	44.81		96.16	○
segment	89.65	89.39		95.11	○
sick	97.03	97.51		99.20	○
sonar	81.33	76.95		79.38	
soybean	89.01	91.50		98.98	○
splice	83.26	90.16	○	86.30	○
vehicle	68.68	68.21		76.96	○
vote	92.90	95.64		96.55	
vowel	67.68	70.61		91.21	○
waveform-5000	74.60	78.00	○	77.90	
zoo	89.18	89.18		91.18	
Average	80.66	82.08		86.58	
w/t/l	-	7/28/1		17/18/1	

○, ● statistically significant upgradation or degradation

REFERENCES

[1] T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill, 1997.

[2] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*, 1st ed. Pearson Education, 2006.

[3] T. T. Cover and P. E. Hart, "Nearest neighbour pattern classification," *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, 1967.

[4] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[5] K. D. Aha, D. and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[6] D. Aha, "Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms," *International Journal of Man-Machine Studies*, vol. 36, pp. 267–287, 1992.

[7] C. G. Atkeson and A. W. Moore, "Locally weighted learning," *Artificial Intelligence Review*, vol. 11, pp. 11–73, 1997.

[8] H. M. Frank, E. and B. Pfahringer, "Locally weighted naive bayes," ser. Proceedings of the Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, 2003, pp. 249–256.

[9] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *Journal of Artificial Intelligence Research*, vol. 6, pp. 1–34, 1997.

[10] C. Stanfill and D. Waltz, "Toward memory-based reasoning," *Communications of the ACM*, vol. 29, pp. 1213–1228, 1986.

[11] R. D. Short and K. Fukunaga, "A new nearest neighbor distance measure," ser. Proceedings of the 5th IEEE Znt. Conf. Pattern Recognition, 1980, pp. 81–86.

[12] R. Short and K. Fukunaga, "The optimal distance measure for nearest neighbour classification," *IEEE Transactions on Information Theory*, vol. 27, pp. 622–627, 1981.

[13] J. P. Myles and D. J. Hand, "The multi-class metric problem in nearest neighbour discrimination rules," *Pattern Recognition*, vol. 23, pp. 1291–1297, 1990.

[14] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, pp. 103–130, 1997.

[15] B. Wang and H. Zhang, "Probability based metrics for locally weighted naive bayes," ser. Proceedings of the 20th Canadian Conference on Artificial Intelligence. Springer, 2007, pp. 180–191.

[16] E. Blanzieri and F. Ricci, "Probability based metrics for nearest neighbor classification and case-based reasoning," ser. Proceedings of the 3rd International Conference on Case-Based Reasoning Research and Development. Springer, 1999, pp. 14–28.

[17] J. G. Cleary and L. E. Trigg, "K*: An instance-based learner using an entropic distance measure," ser. Proceed-

TABLE IV.
EXPERIMENTAL RESULTS FOR OM VERSUS THE IMPROVED OM BY ATTRIBUTE SELECTION: CLASSIFICATION ACCURACY.

Dataset	OM	Wrapper		Filter	
anneal	95.88	90.98	•	87.63	•
anneal.ORIG	84.41	80.29	•	81.73	
audiology	58.79	61.96		64.58	
autos	62.52	58.60		59.60	
balance-scale	83.84	83.84		83.84	
breast-cancer	73.09	73.45		75.87	
breast-w	93.99	94.56		94.42	
colic	83.13	85.03		81.52	
colic.ORIG	69.82	69.55		75.01	
credit-a	86.09	85.51		85.51	
credit-g	71.90	70.30		71.10	
diabetes	69.02	73.56	○	72.40	
glass	57.06	49.05		53.38	
heart-c	81.09	79.16		81.80	
heart-h	82.02	81.06		83.07	
heart-statlog	82.22	74.81		82.59	
hepatitis	84.50	81.83		80.00	
hypothyroid	93.08	93.32		93.40	
ionosphere	89.74	88.03		83.48	•
iris	93.33	94.00		92.67	
kr-vs-kp	95.06	94.09		90.43	•
labor	85.67	79.00		79.33	
letter	71.22	71.70		68.75	•
lymph	80.86	76.24		82.90	
mushroom	99.75	98.28	•	97.60	•
primary-tumor	42.47	35.70		43.63	
segment	89.65	90.00		87.84	
sick	97.03	97.64		96.55	
sonar	81.33	74.48		71.17	•
soybean	89.01	85.63		86.36	
splice	83.26	87.96	○	88.09	○
vehicle	68.68	66.32		61.59	•
vote	92.90	95.18		95.64	
vowel	67.68	61.01		50.40	•
waveform-5000	74.60	76.50		80.00	○
zoo	89.18	80.27	•	86.27	
Average	80.66	78.86		79.17	
w/t/l	-	2/30/4		2/26/8	

○, • statistically significant upgradation or degradation

TABLE V.
EXPERIMENTAL RESULTS FOR OM VERSUS THE IMPROVED OM BY ATTRIBUTE SELECTION: TEST TIME.

Dataset	OM	Wrapper		Filter	
anneal	90.80	40.60	•	21.80	•
anneal.ORIG	104.70	78.10		18.60	•
audiology	10.80	0.00	•	0.00	•
autos	4.70	3.20		3.20	
balance-scale	6.10	10.80		7.70	
breast-cancer	1.60	3.10		1.60	
breast-w	17.00	12.30		14.10	
colic	9.40	1.50		3.10	
colic.ORIG	11.10	6.20		4.70	
credit-a	21.70	20.20		21.80	
credit-g	96.70	25.10	•	17.10	•
diabetes	18.50	7.90	•	9.50	
glass	1.50	1.60		1.60	
heart-c	6.30	3.20		3.10	
heart-h	0.00	1.60		0.00	
heart-statlog	4.80	1.60		4.70	
hepatitis	1.50	1.60		1.50	
hypothyroid	1362.80	1740.60		2596.90	○
ionosphere	15.60	1.60	•	1.50	•
iris	0.00	1.60		0.00	
kr-vs-kp	1048.50	476.70	•	549.90	•
labor	0.00	0.00		0.00	
letter	756.20	562.50	•	432.70	•
lymph	3.10	0.00		0.00	
mushroom	179.70	56.50	•	104.50	•
primary-tumor	6.20	0.00		1.60	
segment	296.90	140.70	•	93.80	•
sick	1159.50	540.70	•	2486.20	○
sonar	6.30	1.60		1.60	
soybean	49.90	23.50	•	31.10	•
splice	1739.10	176.40	•	206.20	•
vehicle	35.90	14.10	•	9.40	•
vote	7.90	4.80		4.80	
vowel	34.40	21.90		11.00	•
waveform-5000	112.40	25.20	•	39.00	•
zoo	0.00	0.00		0.00	
Average	200.60	111.31		186.23	
w/t/l	-	0/22/14		2/20/14	

○, • statistically significant upgradation or degradation

ings of the 12th International Machine Learning Conference. Morgan Kaufmann, 1995, pp. 108–114.

[18] T. Daniel and T. E. Daniel, "Making the nearest neighbor meaningful," ser. Proceedings of SIAM WOrkshop on Clustering High Dimensional Data and its Applications, 2002.

[19] K. G. Han, E. H. and V. Kumar, "Text categorization using weight adjusted k-nearest neighbor classification," University of Minnesota, Tech. Rep., 1999, dep. Comput. Sci.

[20] Z. H. Jiang, L. and Z. Cai, "Dynamic k-nearest-neighbor naive bayes with attribute weighted," ser. Proceedings of the 3rd International Conference on Fuzzy Systems and Knowledge Discovery. Springer, 2006, pp. 365–368.

[21] W. D. C. Z. J. S. Jiang, L. and X. Yan, "Scaling up the accuracy of k-nearest-neighbor classifiers: A naive-bayes hybrid," *International Journal of Computers and Applications*, vol. 31, pp. 36–43, 2009.

[22] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," ser. In Research Issues on Data Mining and Knowledge Discovery, 1997, pp. 1–8.

[23] M. A. Hall, "A decision tree-based attribute weighting filter for naive bayes," *Knowledge-Based Systems*, vol. 20, pp. 120–126, 2007.

[24] D. Wettschereck and D. W. Aha, "Weighting features," ser. Proceedings of the First International Conference on Case-Based Reasoning Research and Development. Springer, 1995, pp. 347–358.

[25] L. P. Kohavi, R. and Y. Yun, "The utility of feature weighting in nearest-neighbor algorithms," ser. Proceedings of the 9th European Conf. on Machine Learning. Springer, 1997, pp. 85–92.

[26] P. Langley and S. Sage, "Induction of selective bayesian classifiers," ser. Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, 1994, pp. 339–406.

[27] G. J. R. R. Beyer, K. and U. Shaft, "When is."

[28] A. C. Hinneburg, A. and D. Keim, "What is the nearest neighbor in high dimensional spaces," ser. Proceedings of the 26th International Conference on Very Large Data Bases, 2000, pp. 506–515.

[29] R. John. G, Kohavi and K. Perfleg, "Irrelevant features and the subset selection problem," ser. Proceedings of the eleventh international conferencce on machine learning. Morgan Kaufmann, 1994, pp. 121–129.

[30] G. John and R. Kohavi, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[31] J. Loughrey and P. Cunningham, "Overfitting in wrapper-based feature subset selection: the harder you try the worse it gets," ser. Proceedings of International Conference on Innovative Techniques and Applications of Artificial Intelligence, 2004, pp. 33–43.

[32] H. Almuallim and T. G. Dietterich, "Learning boolean

TABLE VI.
EXPERIMENTAL RESULTS FOR OM VERSUS THE IMPROVED OM BY ATTRIBUTE SELECTION: THE SIZE OF DIMENSIONALITY.

Data set	OM	Wrapper	Filter
anneal	38	5.7	5.7
anneal.ORIG	38	1.6	3.8
audiology	69	3.2	5.5
autos	25	5	6.8
balance-scale	4	4	4
breast-cancer	9	1.8	3.9
breast-w	9	4.3	8.6
colic	22	5.2	1
colic.ORIG	26	1.7	2.7
credit-a	15	1	1
credit-g	20	4.5	2.8
diabetes	8	1	3
glass	9	2.8	5.3
heart-c	13	4.1	5.7
heart-h	13	1.2	3
heart-statlog	13	4.5	5.9
hepatitis	19	2.9	9.4
hypothyroid	29	2.5	2
ionosphere	34	7	4.1
iris	4	3.8	2
kr-vs-kp	36	4	3
labor	16	2.1	3.7
letter	16	12.1	9
lymph	18	4.5	8.4
mushroom	22	2.9	1
primary-tumor	17	4.8	11.7
segment	19	7.8	5
sick	29	3.4	1
sonar	60	1.1	7.4
soybean	35	12.7	21.9
splice	60	4.2	6
vehicle	18	5.8	3.8
vote	16	1.1	1
vowel	13	6.4	2
waveform-5000	40	8.9	14
zoo	16	4.2	7.8
Average	23.56	4.27	5.36

concepts in the presence of many irrelevant features," *Artificial Intelligence*, vol. 69, pp. 279–305, 1994.

[33] K. Kira and L. A. Rendell, "A practical approach to feature selection," ser. Proceedings of the Ninth International Conference. Morgan Kaufmann, 1992, pp. 249–256.

[34] M. Scherf and W. Brauer, "Feature selection by means of a feature weighting approach," Forschungsberichte Kunstliche Intelligenz, Institut fur Informatik, Technische Universitat Munchen, Tech. Rep., 1997.

[35] M. A. Hall, "Correlation-based feature subset selection for machine learning," University of Waikato, Tech. Rep., 1998, dep. Comput. Sci.

[36] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," ser. Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 359–366.

[37] Y. Lei and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[38] L. F. Peng, H. and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE trasaction on pattern analysis and machine intelegence*, vol. 27, pp. 1226–1238, 2005.

[39] C. Cardie, "Using decision trees to improve cased-based learning," ser. Proceedings of the tenth international conferencce on machine learning. Morgan Kaufmann, 1993, pp. 25–32.

[40] C. A. Ratanamahatana and D. Gunopulos, "Scaling up the naive bayesian classifier: Using decision trees for feature selection," ser. Proceedings of the Workshop on Data Cleaning and Preprocessing, 2002.

[41] M. C. A. Mitra, P. and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 301–312, 2002.

[42] L. Huan and R. Setiono, "Dimensionality reduction via discretization," *Knowledge-Based Systems*, vol. 9, pp. 67–72, 1996.

[43] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, pp. 239–281, 2003.

Chaoqun Li is currently a Ph.D. candidate at China University of Geosciences (Wuhan). Her research interests include data mining and machine learning.

Hongwei Li is currently the doctoral supervisor of Chaoqun Li and a professor in Department of Mathematics at China University of Geosciences (Wuhan).