Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

Performance evaluation of classification algorithms by *k*-fold and leave-one-out cross validation

Tzu-Tsung Wong*

Institute of Information Management National Cheng Kung University 1, Ta-Sheuh Road, Tainan City 701, Taiwan, ROC

ARTICLE INFO

Article history:
Received 6 November 2014
Received in revised form
4 February 2015
Accepted 8 March 2015

Keywords:
Classification
Independence
k-Fold cross validation
Leave-one-out cross validation
Sampling distribution

ABSTRACT

Classification is an essential task for predicting the class values of new instances. Both *k*-fold and leave-one-out cross validation are very popular for evaluating the performance of classification algorithms. Many data mining literatures introduce the operations for these two kinds of cross validation and the statistical methods that can be used to analyze the resulting accuracies of algorithms, while those contents are generally not all consistent. Analysts can therefore be confused in performing a cross validation procedure. In this paper, the independence assumptions in cross validation are introduced, and the circumstances that satisfy the assumptions are also addressed. The independence assumptions are then used to derive the sampling distributions of the point estimators for *k*-fold and leave-one-out cross validation. The cross validation procedure to have such sampling distributions is discussed to provide new insights in evaluating the performance of classification algorithms.

1. Introduction

Data mining is an emerged technique for automatically processing the huge amount of data stored in computers, and classification is an essential task in data mining for assigning the class values of new instances. Two popular approaches for evaluating the performance of a classification algorithm on a data set are *k*-fold and leave-one-out cross validation. When the amount of data is large, *k*-fold cross validation should be employed to estimate the accuracy of the model induced from a classification algorithm, because the accuracy resulting from the training data of the model is generally too optimistic [1]. Leave-one-out cross validation is a special case of *k*-fold cross validation, in which the number of folds equals the number of instances. When the number of instances either in a data set or for a class value is small, such as gene microarray data and gene sequence data, leave-one-out cross validation should be adopted to obtain a reliable accuracy estimate for a classification algorithm. Unlike leave-one-out cross validation, there is a randomness mechanism in *k*-fold cross validation such that the mean accuracy resulting from *k*-fold cross validation on a data set is not a constant.

Bias and variance are two main measures for investigating the impact of this randomness mechanism, where bias represents the expected difference between an accuracy estimate and actual accuracy, and variance represents the variability of an accuracy estimate. The concept of bias and variance is employed to argue why simple classification algorithms such as naïve Bayesian classifiers and k-nearest neighbors can achieve competitive performance [2], and Bengio and Grandvalet [3] showed that the unbiased estimator of the variance of *k*-fold cross validation does not exist. The bias of accuracy estimate will be smaller when the number of folds is either five or ten [4].

The following four factors can affect an accuracy estimate obtained by *k*-fold cross validation.

- The number of folds.
- The number of instances in a fold.
- The level of averaging.
- The repetition of *k*-fold cross validation.

Introducing the ways of executing *k*-fold and leave-one-out cross validation are almost necessary in every book for data mining. Table 1 summarizes the context of the two approaches in six books, including whether the way to make statistical inference for the results obtained by leave-one-out cross validation is introduced, where a hyphen denotes that a book does not explicitly describe the item corresponding to a column. The four factors for executing *k*-fold cross validation are not all the same in every pair of books in Table 1. This implies that not only the randomness mechanism for dividing instances into folds, but also the settings of the four factors can affect the accuracy estimate obtained by *k*-fold cross validation.

Every book in Table 1 provides the ways to derive confidence interval or to perform hypothesis testing for the results obtained

* Tel.: +886 6 2757575x53722; fax: +886 6 2362162.
E-mail address: tzutsung@mail.ncku.edu.tw

http://dx.doi.org/10.1016/j.patcog.2015.03.009
0031-3203/

Table 1
A summary for *k*-fold and leave-one-out cross validation.

	k-Fold cross validation				Leave-one-out cross validation
	No. of folds	Fold size	Averaging level	Repetition	
Alpaydin [5]	10 or 30	$np \geq 5$ and $n(1-p) \geq 5$[a]	Fold	Suggested	–
Han et al. [6]	10	–	Data set	Suggested	–
Kantardzic [7]	–	–	Fold	–	–
Mitchell [8]	–	30	Fold	–	–
Tan et al. [9]	Large	–	Data set	–	–
Witten et al. [1]	10	100	Fold	Suggested	–

[a] n and p represent the number of instances in a fold and actual accuracy, respectively.

by *k*-fold cross validation. In leave-one-out cross validation, every fold has only one instance, and hence random partition is not necessary. The ways of making statistical inference on the results obtained by *k*-fold cross validation cannot therefore be applied to analyze those obtained by leave-one-out cross validation. However, none of the six books given in Table 1 introduces statistical methods specifically suitable for evaluating the performance of a classification algorithm by leave-one-out cross validation.

For any population parameter, a necessary step for deriving confidence interval or performing hypothesis testing is to know the sampling distribution of its corresponding point estimator. In this paper, we will derive the sampling distributions of the point estimators for *k*-fold and leave-one-out cross validation. The sampling distributions will be used to analyze the proper settings of the four factors in performing *k*-fold cross validation, and to design statistical inference methods for leave-one-out cross validation.

This paper is organized as follows. Section 2 briefly introduces the definition of sampling distribution and the central limit theorem. A sample is generally assumed to be collected by simple random sampling for deriving the sampling distribution of a point estimator. The observations in a random sample need to be independent and come from the same population. Section 3 therefore discusses the independence properties about the testing results obtained by *k*-fold and leave-one-out cross validation. The sampling distributions for *k*-fold and leave-one-out cross validation are derived in Sections 4 and 5, respectively. The conclusions and future directions for research are summarized in Section 6.

2. Sampling distributions

Accuracy is a critical measure for evaluating the performance of a classification algorithm. When all instances in a data set have the same weight, the accuracy of a classification algorithm on a data set is defined as the number of instances predicted correctly over the total number of instances. In this case, accuracy is considered as a sample proportion that is a special case of a sample mean. Sample mean $\overline{x}$ and sample proportion $\overline{p}$ are point estimators of population mean μ and population proportion p, respectively, and the probability distributions governing point estimators are called sampling distributions. This section will briefly introduce the sampling distributions of $\overline{x}$ and $\overline{p}$. The sampling distribution of the sample mean can be approximated by a normal distribution as the same size becomes large.

The usefulness of the central limit theorem is that regardless of the probability distribution governing a population, sample mean can be assumed to have a normal distribution when sample size is large. Since sample proportion is a special case of sample mean, the central limit theorem can be applied to sample proportion as well. If a population follows a normal distribution, then the sampling distribution of a sample mean calculated from independent observations will be normally distributed for any sample size, because the sum of two independent normal random variables is also normally distributed [10]. When the probability distribution governing a population is unknown, the sampling distribution will be approximately normal only when the central limit theorem can be applied. In this case, the criteria to determine whether a sample size n is large are necessary.

Let x_i be the ith observation in a sample. Many statistics literatures suggest that if the probability distribution for the population is not highly skewed, a sample is large enough to assume a normal distribution for the sample mean if sample size $n \geq 30$. When the population proportion p is very close to zero or one, $n \geq 30$ is no longer an appropriate criterion for applying the central limit theorem. The criterion about large sample is therefore revised as $np \geq 5$ and $n(1-p) \geq 5$ for sample proportion [11,12], which are derived from the normal approximation of binomial distribution. For instance, if $p=0.9$, a sample with size $n=40$ is not large enough to assume that $\overline{p}$ follows a normal distribution because $n(1-p)=4<5$. Note that np and $n(1-p)$ represent the expected number of successes and failures, respectively in a sample with size n. This means that the resulting accuracy of a classification algorithm can be assumed to follow a normal distribution if the expected numbers of correct and wrong predictions on n instances are both not less than five.

The procedure for obtaining a simple random sample with size n from a finite population is to control that every possible sample of size n has the same probability to be chosen. When the size of a population is infinite, the number of possible samples of size n will also be infinite. A random sample of size n from an infinite population must therefore satisfy two conditions: each observation is selected independently and comes from the same population [12]. To ensure that the predictions of two instances are independent, they must be independently drawn from the same population. Let $x_i=1$ if the prediction of instance i by a classification algorithm is correct, and $x_i=0$ otherwise. Then we have $E(x_i)=p$ and $\mathrm{Var}(x_i)=p(1-p)$, where p is the prediction accuracy of the model induced by the algorithm on the population. The sampling distribution of $\overline{p}=\sum_{i=1}^{n} x_i/n$ can be approximated by $\mathrm{N}(p, p(1-p)/n)$ by the central limit theorem if $np \geq 5$ and $n(1-p) \geq 5$.

Since the population proportion p is generally unknown, it is not easy to determine whether conditions $np \geq 5$ and $n(1-p) \geq 5$ hold for a random sample with n instances. Hence, in identifying whether a random sample with size n is large or not, we will replace p by $\overline{p}$ in the two conditions, because $\overline{p}$ is an unbiased estimator of p. In the remainder of this paper, a random sample with n instances is therefore considered to be large if the numbers of correct and wrong predictions are both not less than five, and they are called large-sample conditions. For instance, if the number of correct predictions for a random sample with 50 instances is 42, then the resulting accuracy can be assumed to be normally distributed. On the contrary, if the number of correct predictions for this sample is 47, then it is inappropriate to assume that the resulting accuracy follows a normal distribution, because the number of wrong predictions is only three.

3. Independence

Most statistical inference techniques need data to be collected by simple random sampling. As described in the previous section, when population size is infinite, two necessary conditions for simple random sampling are: every observation comes from the same population, and all observations are collected independently.

Please cite this article as: T.-T. Wong, Performance evaluation of classification algorithms by *k*-fold and leave-one-out cross validation, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.03.009

The instances in a data set, called a random sample, are generally assumed to come from the same population. Otherwise, they should not be in the same data set for learning. It is therefore reasonable to assume that the instances in a data set are all governed by the probability distribution for a population. When data are collected by simple random sampling, every two instances in a data set are considered to be independent. This section will discuss the impact of independence assumptions on the point estimators obtained by cross validation.

The purpose of classification is to find a model from training data such that the model can have a correct prediction on the class value for most new instances. Let A, R, and e represent classification algorithm, training data set, and new instance, respectively. Then the model learned from training data R by classification algorithm A can be represented as $M_{A,R}$, and hence the class value predicted by this model for instance e can be denoted by $M_{A,R}(e)$. Let $c(e)$ be the actual class value of instance e, and let p_A be the actual probability of correct prediction of classification algorithm A on the population; i.e., $p\{M_{A,R}(e)=c(e)\}=p_A$. This expression implies that whether a prediction is correct or not depends on classification algorithm, training data, and new instance.

Definition 1. Let the instances for training and testing be independent.

(a) Instance-independence assumption: For any two independent instances e_1 and e_2, $M_{A,R}(e_1)=c(e_1)$ is independent of $M_{B,R'}(e_2)=c(e_2)$ for any A, B, R, and R'.
(b) Scheme-independence assumption: For any two different classification methods A and B, $M_{A,R}(e_1)=c(e_1)$ is independent of $M_{B,R'}(e_2)=c(e_2)$ for any R, R', e_1, and e_2.

In cross validation, an instance in R cannot be for testing. It is therefore reasonable to assume that any new instance is independent of training data set R. Two new instances e_1 and e_2 are independent when they are collected by simple random sampling. The predictions of e_1 and e_2 are therefore independent regardless of classification algorithms and training data for generating prediction models. This means that the instance-independence assumption is always true for both k-fold and leave-one-out cross validation. However, the conditions under which the scheme-independence assumption is true will be more complicate.

Every classification algorithm has its own mechanism in learning a model from training data. It seems that assuming that $M_{A,R}(e)$ is independent of $M_{B,R}(e)$ for two different classification algorithms A and B should be reasonable. Let A be the algorithm for finding a fully grown decision tree by the gain ratio from R, and let B be the algorithm for finding a fully grown decision tree by the gain ratio from R and pruning this tree by a measure. The only difference between algorithms A and B is that B has a mechanism to prune the fully grown tree. Since the fully grown trees found by algorithms A and B must be identical, it is therefore inappropriate to assume that $M_{A,R}(e)$ is independent of $M_{B,R}(e)$ in this case.

As addressed by Mitchell [8], a classification algorithm must have inductive bias to predict class values for new instances, and the inductive bias can be language bias, search bias, or both. The models considered in the learning mechanism of a classification algorithm form a model space. A classification algorithm will have no language bias if all possible models have been included in its model space. The search bias describes whether a classification algorithm has used a measure to prefer a model over another. If the model spaces of two classification algorithms have nonequivalent representation, then the scheme-independence assumption must be true. For instance, a model in decision tree induction is a decision tree, and a model in support vector machine is a hyperplane. Since the two algorithms have nonequivalent representation of a model, the scheme-independence assumption is true in comparing their performance. Note that a decision tree can be represented as a set of classification rules. The model spaces of decision tree induction and sequential covering algorithm will have common models.

When two classification algorithms have the same model space, they can still be scheme-independent if their preference on models are independent. For instance, the growing measure in decision tree induction can be gain ratio or gini index. Since the two measures set different preference on models, the two models found by gain ratio and gini index can be assumed to be independent. In summary, if the model spaces of two classification algorithms do not have equivalent representation, then the scheme-independence assumption is true regardless of their search bias. When the model spaces of two algorithms have common models, then the scheme-independence assumption can be true only when they have independent search bias. This guideline can also be applied to analyze whether the scheme-independence assumption is true for discretization and feature selection, two popular tasks in data preprocessing.

Discretization transforms continuous attributes into discrete ones. The two main operations of discretization is to determine the number of intervals and the boundaries of every interval. If two discretization methods perform the two operations independently, then the models represented by the discretized continuous attributes can be assumed to be nonequivalent. In this case, the scheme-independence assumption is true even the classification algorithms for evaluating the two discretization methods are the same. For instance, the model spaces formed by the discrete attributes resulting from the equal-width discretization and the entropy-based discretization are nonequivalent.

Feature selection is a tool to remove redundant and irrelevant attributes for classification. Suppose that the classification algorithms for evaluating two feature selection methods are the same. If the intersection of the attribute subsets chosen by the two feature selection methods is not empty, then the model spaces resulting from the two attribute subsets will have common models. The scheme-independence assumption is therefore false in this case.

In leave-one-out cross validation, every instance is in turn used to test the model induced from the other instances. Thus, the instance-independence assumption guarantees that every prediction in leave-one-out cross validation is independent of each other. The scheme-independence assumption indicates that $M_{A,R}(e)=c(e)$ is independent of $M_{B,R}(e)=c(e)$. This assumption provides a base for comparing the performance of two classification algorithms by cross validation.

Though training data can affect the result of a prediction, they cannot be used to determine whether two predictions are independent without considering testing instances and classification algorithms. When two training data sets R and R' are both collected by simple random sampling, R and R' are independent. Since they are governed by the same probability distribution for the population, it is possible that a classification algorithm will induce the same model from R and R'. In this case, $M_{A,R}(e)$ and $M_{A,R'}(e)$ will be the same. It is therefore inappropriate to assume that training data sets are independent in cross validation.

Proposition 1. *The k accuracies resulting from k-fold cross validation are independent.*

Proof. Let a data set D be divided into folds $F_1, F_2, \ldots, F_k$ such that $F_i \cap F_j = \varnothing$ for any $i \neq j$. In evaluating the performance of classification algorithm A, the accuracy of fold F_j is calculated as $\Sigma_{e \in F_j} I(M_{A,D\setminus F_j}(e) = c(e))/|F_j|$, where $|F_j|$ is the number of instances in F_j, and $I(Y=y)$ is an indicator function that has value one when condition $Y=y$ holds, and zero otherwise. Since $F_i \cap F_j = \varnothing$ for any $i \neq j$, by the instance-independence assumption, $\Sigma_{e \in F_i} I(M_{A,D\setminus F_i}(e) = c(e))/|F_i|$ and $\Sigma_{e \in F_j} I(M_{A,D\setminus F_j}(e) = c(e))/|F_j|$ are independent.

Please cite this article as: T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.03.009

Proposition 1 indicates that the k accuracies obtained by k-fold cross validation are all independent. Since they are generated by the same procedure, the k accuracies are considered to come from the same population. The k accuracies resulting from k-fold cross validation can therefore be observations in a sample collected by simple random sampling for estimating the actual accuracy of a classification algorithm.

Some studies randomly choose a specific proportion of instances from a data set to be testing data, and the remaining instances are for training. This procedure can be repeated to obtain several accuracies for evaluating the performance of a classification algorithm. The mean value of the accuracies is an unbiased point estimator of the actual accuracy of the classification algorithm. However, since the accuracies are not independent, it will be very difficult to derive a sampling distribution for statistical inference.

4. *k*-Fold cross validation

A popular procedure for estimating the performance of a classification algorithm or comparing the performance between two classification algorithms on a data set is k-fold cross validation. This procedure randomly divides a data set into k disjoint folds with approximately equal size, and each fold is in turn used to test the model induced from the other $k-1$ folds by a classification algorithm. The performance of the classification algorithm is evaluated by the average of the k accuracies resulting from k-fold cross validation, and hence the level of averaging is assumed to be at fold. This section will present the sampling distributions of the point estimators for k-fold cross validation, and discuss the appropriate way of its application. All folds are assumed to contain the same number of instances except explicitly specified.

4.1. Single algorithm

Let a data set D be divided into disjoint folds $F_1, F_2, \ldots, F_k$, and let $|D|=n$ and $|F_j|=m$ be the number of instances in D and F_j, respectively; i.e., $n=km$. Furthermore, let the number of correct predictions on F_j by a classification algorithm A be r_j, and let the prediction accuracy of A on the whole population corresponding to data set D be p. The following theorem presents the necessary conditions for the mean accuracy resulting from k-fold cross to be approximately normally distributed.

Theorem 1. *If $r_j \geq 5$ and $m-r_j \geq 5$ for $j=1, 2, \ldots, k$, then the sampling distribution of the resulting mean accuracy $\overline{p}=\sum_{j=1}^{k} p_j/k$ can be approximated by a normal distribution with mean p and variance $p(1-p)/n$, where $p_j = r_j/m$ for $j = 1, 2, \ldots, k$*

Proof. According to the discussion given in Section 2, if $r_j \geq 5$ and $m-r_j \geq 5$, then by the central limit theorem, $p_j=r_j/m$ can be assumed to follow a normal distribution with mean p and variance $p(1-p)/m$. Since every pair of instances in D are independent, and $F_i \cap F_j=\varnothing$ for any $i \neq j$, by Proposition 1, accuracies p_1 through p_k are independent and identically distributed random variables. Since the sum of independent random variables governed by normal distributions is also normally distributed, the sampling distribution of $\overline{p}=\sum_{j=1}^{k} p_j/k$ can be approximated by a normal distribution. We also have

$$E(\overline{p})=\frac{\sum_{j=1}^{k} E(p_j)}{k}=p$$

and

$$\mathrm{Var}(\overline{p})=\frac{\sum_{j=1}^{k} \mathrm{Var}(p_j)}{k^2}=\frac{p(1-p)}{n}.$$

Theorem 1 shows that if the testing results of all folds satisfy the large-sample conditions, the sampling distribution of $\overline{p}$ can be assumed to be a normal distribution regardless of the number of folds. However, when either $r_j < 5$ or $m-r_j < 5$ for fold F_j, p_j should not be used to calculate $\overline{p}$. Otherwise, the sampling distribution of $\overline{p}$ will not be approximately normal for estimating the accuracy of the classification algorithm.

4.2. Two algorithms

There are two ways to compare the prediction accuracies of two classification algorithms by k-fold cross validation. When it is possible to evaluate the two algorithms by the same data in each iteration, the matched sample approach is more suitable for this purpose. If the testing data for two algorithms in an iteration are different, then we should adopt the independent sample approach to compare their accuracies.

Let p_A and p_B be the prediction accuracies of classification algorithms A and B, respectively on the population corresponding to a data set D. When D is randomly divided into disjoint folds F_1, F_2, …, F_k, both A and B are trained by the instances in $D \backslash F_j$ and tested by F_j in the jth iteration. Assume as before that $|D|=n$ and $|F_j|=m$ for $j=1, 2, \ldots, k$. Let r_{ij} be the number of instances in F_j correctly classified by algorithm i for $i=A, B$ and $j=1, 2, \ldots, k$. In this matched sample case, the point estimator for identifying whether it is appropriate to assume $p_A=p_B$ is calculated as $\overline{d}=\sum_{j=1}^{k} d_j/k$, where $d_j=(r_{Aj}-r_{Bj})/m$ for $j=1, 2, \ldots, k$.

Theorem 2. *If $r_{ij} \geq 5$ and $m-r_{ij} \geq 5$ for $i=A, B$ and $j=1, 2, \ldots, k$, then the sampling distribution of the point estimator $\overline{d}=\sum_{j=1}^{k} d_j/k$ can be approximated by a normal distribution with mean p_A-p_B when the scheme-independence assumption is satisfied.*

Proof. Since $r_{ij} \geq 5$ and $m-r_{ij} \geq 5$ for $i=A, B$ and $j=1, 2, \ldots, k$, r_{ij}/m can be assumed to follow a normal distribution with mean p_i. When the scheme-independence assumption is satisfied, point estimators r_{Aj}/m and r_{Bj}/m are independent. Hence, the sampling distribution of $d_j=r_{Aj}/m-r_{Bj}/m$ can be approximated by a normal distribution. We also have

$$E(\overline{d})=E\left(\sum_{j=1}^{k} d_j/k\right)=\sum_{j=1}^{k} E(d_j)/k=p_A-p_B.$$

When classification algorithms A and B are independent, it can be shown that

$$\mathrm{Var}(d_j)=\frac{p_A(1-p_A)}{m}+\frac{p_B(1-p_B)}{m}.$$

Note that $\overline{d}$ is a sample mean instead of a sample proportion, and that both p_A and p_B are unknown. The variance of sample $\{d_1, d_2, \ldots, d_k\}$ is calculated as $s_d^2=\sum_{j=1}^{k}(d_j-\overline{d})^2/(k-1)$ that is an estimate of $\mathrm{Var}(d_j)$, and t value will be the test statistic in this case. When the null hypothesis is H_0: $p_A-p_B=0$ with significance level α, the test statistic is calculated as $t=\overline{d}/(s_d/\sqrt{k})$ with $k-1$ degrees of freedom. The two classification algorithms A and B will have significantly different accuracy if the p-value corresponding to the t value is less than α. If the conditions specified in Theorem 2 hold, this matched sample approach can be applied for any value of k.

When the testing data at each iteration for the two algorithms are different, the independent sample approach can be used to

Please cite this article as: T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.03.009

derive the sampling distribution for comparing their performance. Let $p_{Aj}=r_{Aj}/m$ be the accuracy of A at iteration j for $j=1, 2, \ldots, k$, and similarly for p_{Bj} for $j=1, 2, \ldots, q$. Then $\overline{p}_A=\sum_{j=1}^{k} p_{Aj}/k$ and $\overline{p}_B=\sum_{j=1}^{q} p_{Bj}/q$ are unbiased estimators of p_A and p_B, respectively. Hence, the point estimator of p_A-p_B is $\overline{p}_A-\overline{p}_B$ in this case.

Theorem 3. *If the testing results of the $k+q$ folds for evaluating the performance of algorithms A and B are all satisfy the large-sample conditions, then the sampling distribution of the point estimator $\overline{p}_A-\overline{p}_B$ can be approximated by a normal distribution with mean p_A-p_B and variance $(p_A(1-p_A)/n)+(p_B(1-p_B)/n)$ when the scheme-independence assumption is true.*

Proof. Since the numbers of correct and wrong predictions in each fold are not less five, by Theorem 1, $\overline{p}_i$ approximately follows a normal distribution with mean p_i and variance $p_i(1-p_i)/n$ for $i=A, B$. When the scheme-independence assumption is true, the prediction of an instance by a model induced by algorithm A is independent of the prediction of the same instance by a model learned by algorithm B. This implies that $\overline{p}_A$ and $\overline{p}_B$ are independent, and hence the sampling distribution of $\overline{p}_A-\overline{p}_B$ can be assumed to be normally distributed. We also have $E(\overline{p}_A-\overline{p}_B)=p_A-p_B$ and

$$\mathrm{Var}(\overline{p}_A-\overline{p}_B)=\mathrm{Var}(\overline{p}_A)+\mathrm{Var}(\overline{p}_B)=\frac{p_A(1-p_A)}{n}+\frac{p_B(1-p_B)}{n}.$$

When the null hypothesis is H_0: $p_A-p_B=0$, the two samples for estimating $\overline{p}_A$ and $\overline{p}_B$ are pooled together to calculate a more reliable estimate for p_A and p_B as $\overline{p}=(\overline{p}_A+\overline{p}_B)/2$. The test statistic is therefore calculated as $z=(\overline{p}_A-\overline{p}_B)/\sqrt{2\overline{p}(1-\overline{p})/n}$.

There is another way to calculate a test statistic for comparing the performance of two classification algorithms A and B. Compute $\overline{p}_A=\sum_{i=1}^{k} p_{Ai}/k$ and $\overline{p}_B=\sum_{j=1}^{q} p_{Bj}/q$ as before. If the numbers of correct and wrong predictions in every fold are not less than five, then $\overline{p}_A-\overline{p}_B$ can be assumed to have a normal distribution with mean p_A-p_B, and the variance of this distribution can be estimated as $(s_A^2/k)+(s_B^2/q)$, where $s_A^2=(\Sigma_{i=1}^{k}(p_{Ai}-\overline{p}_A)^2/(k-1)$ and $s_B^2=(\Sigma_{j=1}^{q}(p_{Bj}-\overline{p}_B)^2/(q-1)$. The test statistics is therefore $t=(\overline{p}_A-\overline{p}_B)/(s_A^2/k+s_B^2/q)$. This way employs the hypothesis testing for comparing two population means by the independent sample approach. As discussed in the previous paragraph, the test statistic for this purpose can be z value that provides a more precise information to determine the hypothesis testing result.

4.3. Discussion

As described in Section 1, there are four factors that can affect the results obtained by k-fold cross validation. The sampling distributions derived in the previous two subsections will be used to investigate the appropriate settings of the four factors in performing k-fold cross validation.

4.3.1. The number of folds

Theorem 1 shows that the sampling distribution of the mean accuracy obtained by k-fold cross validation is independent of the number of folds k. This is reasonable because the sampling distributions for various values of k are all derived from the same testing instances. When k is large, the number of training instances becomes large in each iteration, while the computational cost of k-fold cross validation will be high, and the number of instances in a fold will be small. This implies that the testing results of the instances in a fold have a larger chance to violate the large-sample conditions. In this case, the variance of the sampling distribution estimated by $\overline{p}(1-\overline{p})/n$ will be larger, because the number of instances in the fold that does not satisfy the large-sample conditions cannot be added into n. When the testing results of all folds satisfy the large-sample conditions, the number of folds can therefore be as large as possible.

4.3.2. The number of instances in a fold

Most literatures define the first step of k-fold cross validation as: randomly partition a data set into k disjoint folds with approximately equal size. The way to perform this operation is not unique. The followings are three possible ways to divide a data set with 203 instances into $k=5$ folds.

(1) Each instance is independently assigned to fold j that equals the smallest integer larger than or equal to $5u$, where u is a random number larger than zero. For instance, an instance will be assigned to the second fold when $u=0.26$. Let the number of instances in the five folds be 38, 41, 44, 37, and 43, respectively.
(2) Generate a random number for each instance, and assign the random numbers into ascending order. Then divide the instances into five folds according to the sorted random numbers, and the number of instances in the five folds is 40, 40, 40, 40, and 43, respectively.
(3) The process of generating and sorting random numbers is the same as in approach (2), while the number of instances in the five folds is 40, 40, 41, 41, and 41, respectively.

The mean accuracies resulting from the three approaches are generally different, and which one should be adopted to divide this data set into five folds?

Let m_j be the number of instances in fold j. The p_j for $j=1, 2, \ldots, k$ are all unbiased estimators of actual accuracy p. If they are considered as the observations for estimating p, then they should come from the same population; i.e., they should follow the same normal distribution. This means that they should have, or at least approximately, the same variance. By Theorem 1, p_j can be assumed to have a normal distribution with mean p and variance $p(1\text{-}p)/\mathrm{m_j}$. The m_j for $j=1, 2, \ldots, k$ should therefore be as close to each other as possible. Hence, approach (3) is the most recommended one to divide a data set into folds. When the number of instances n in a data set is large such that n/k is far larger than k, the other two approaches can also be adopted.

The large-sample conditions indicates that it is inappropriate to decide whether a sample is large or not only by the number of instances. For instance, when the classification accuracy of an algorithm on a data set is close to 50%, the testing results of a fold that contains only 20 instances is likely to satisfy the large-sample conditions. On the contrary, when an algorithm has close to 100% prediction accuracy on a data set, a fold containing more than 200 instances may fail to be a large sample.

4.3.3. The level of averaging

The averaging for accuracy estimates can be performed in the level of fold or data set. The level of averaging in deriving sampling distributions is set at fold, and hence the large-sample conditions are checked fold by fold. When an accuracy estimate is calculated by the total number of correct predictions in k folds over the number of instances in a data set, it can be shown that the sampling distribution of this point estimator can be approximated by $N(p, p(1-p)/n)$, the same as the one given in Theorem 1.

Example 1. A data set containing 200 instances is randomly divided into five folds for evaluating the performance of a classification algorithm, and the number of correct predictions in the five folds is 32, 28, 30, 32, and 28, respectively. If the level of

Please cite this article as: T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.03.009

averaging is at data set, the mean and variance of the accuracy estimator for the classification algorithm are calculated as

$\overline{p} = (32+28+30+30+32)/5 = 0.76$

and

$\mathrm{Var}(\overline{p}) = 0.76(1-0.76)/200 = 0.000912.$

The interval with confidence level $1-\alpha$ is $0.76 \pm 0.030 z_{\alpha/2}$. If the level of averaging is at fold, these two values are calculated as

$$\overline{p} = \frac{0.80+0.70+0.75+0.75+0.80}{5} = 0.76$$

and

$$\mathrm{Var}(\overline{p}) = \frac{0.04^2+(-0.06)^2+(-0.01)^2+(-0.01)^2+0.04^2}{5-1} = 0.0007.$$

In this case, the point estimator $\overline{p}$ is a sample mean instead of a sample proportion. Hence, the interval with confidence level $1-\alpha$ is $0.76 \pm t_{\alpha/2}\sqrt{(0.0007/5)} = 0.76 \pm 0.012 t_{\alpha/2}$.

Example 1 shows that when the level of averaging is set at data set, the variance of a point estimator depends only on the sample proportion. In this case, when two classification algorithms evaluated by k-fold cross validation have the same mean accuracy, the variance of the two sample proportions will be the same. We will not be able to know which algorithm has a more stable performance. This will not occur when prediction accuracy is calculated fold by fold. Hence, if the testing results of every fold satisfy the large-sample conditions, the level of averaging should be set at fold.

4.3.4. Repetition

As shown in Table 1, several literatures suggest that k-fold cross validation can be repeatedly performed to obtain several unbiased estimates such that the point estimate of p can be more reliable. For instance, let p_j and p'_j for $j=1, 2, \ldots, k$ be the estimates obtained by the first and the second rounds of k-fold cross validation, respectively. Then $\overline{p}' = \sum_{j=1}^{k}(p_j+p'_j)/2k$ should be a better estimate of p than $\overline{p} = \sum_{j=1}^{k} p_j/k$, because $\overline{p}'$ is an unbiased estimator obtained from a larger sample. Theoretically, $\overline{p}'$ will have a smaller variance than $\overline{p}$.

An interesting result here is that the expected difference between the aggregate point estimator and p will become smaller when the rounds for performing k-fold cross validation becomes larger. Since the number of instances for learning does not increase, what is the new information to make the point estimate more precise? Note that $\overline{p}'$ is a more reliable estimate of p than $\overline{p}$ if the p_j and the p'_j for $j=1, 2, \ldots, k$ are all independent. The same data are used in the first and the second rounds for random partition. An instances in fold i of the first round will be assigned to fold j for some j in the second round. Since the classification algorithm is still the same one, neither the instance-independence nor the scheme-independence assumption holds in this case. This means that p_i and p'_j are not independent. The prediction of the same instance in the first and the second rounds will be positively correlated. For any two random variables X and Y, we have $\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)$. The variance of $X+Y$ increases when X and Y are positively correlated. Since p_i is positively correlated to some of the p'_j, the variance of $\overline{p}'$ actually does not reduce even it is calculated from a larger sample. This explains the myth that repeatedly performing k-fold cross validation can obtain a more reliable estimate of p.

In comparing the performance of two classification algorithms A and B, let $\hat{p}_i = \sum_{j=1}^{k} r_{ij}/n$ for $i=A, B$ when the averaging level is set at data set. As argued previously, it is inappropriate to perform k-fold cross validation repeatedly. The matched sample approach therefore cannot be used to determine whether p_A and p_B are significantly different in this case. The point estimator $\hat{p}_A - \hat{p}_B$ will have the same sampling distribution as $\overline{p}_A - \overline{p}_B$ in the independent sample approach.

In dividing the instances in a data set into independent training and testing data, an instance can play only one role in an iteration. Another approach that also satisfies this requirement is to randomly choose a pre-specified proportion of instances from a data set for testing, and the instances not chosen for testing are for training. If this procedure is executed only once, and the testing results satisfy the large-sample conditions, then we will have a sampling distribution to make statistical inference about prediction accuracy. Since the number of testing instances is generally far less than the number of instances in a data set, the variance of the point estimator resulting from this approach will be larger than that resulting from k-fold cross validation. The procedure to randomly choose testing instances should not be repeated, because the testing sets in two rounds may have common instances. In this case, the testing results of the two rounds will not be independent, and hence they should not be aggregated together for deriving a more reliable point estimate.

5. Leave-one-out cross validation

Many studies adopt leave-one-out cross validation to evaluate the performance of a classification algorithm when the number of instances in a data set or the number of instances for a class value is small. Since the randomness of dividing instances into for training and testing does not exist, the point estimate of accuracy for a given data set is constant. This section will derive the sampling distributions for leave-one-out cross validation to make statistical inference about the mean accuracies of classification algorithms.

5.1. Single algorithm

The prediction of an instance can be either correct or wrong. This means the random variable corresponding to the prediction of an instance follows a Bernoulli distribution with success probability p. When every instance is independent of each other, the number of correct predictions on n instances will follow a binomial distribution with parameters n and p. Let x_i be the random variable corresponding to the prediction of the ith instance. Then $P\{x_i=1\}=p$, and $P\{x_i=0\}=1-p$. Hence, the sampling distribution of the point estimator $\overline{p} = \sum_{i=1}^{n} x_i/n$ is approximated by $N(p, p(1-p)/n)$ when $\sum_{i=1}^{n} x_i \geq 5$ and $n - \sum_{i=1}^{n} x_i \geq 5$.

Note that this sampling distribution is the same as the one obtained in Theorem 1 This is reasonable because leave-one-out cross validation is a special case of k-fold cross validation. Since k-fold cross validation is more efficient, leave-one-out cross validation will be used only when the random partition in k-fold cross validation have a large impact on performance evaluation.

Unlike k-fold cross validation, the sample variance obtained by leave-one-out cross validation is constant. If leave-one-out cross validation is executed several rounds for a data set, every round will have the same resulting sample mean and sample variance. These sample means cannot be pooled together to derive a more reliable point estimate because their variance equals zero. It is therefore helpless to repeatedly perform leave-one-out cross validation for obtaining a more reliable point estimate. This provides another explanation about why k-fold cross validation should be executed only once.

Please cite this article as: T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.03.009

5.2. Two algorithms

Similar to k-fold cross validation, both the independent and the matched sample approaches can be used to compare the performance between two classification algorithms in leave-one-out cross validation. Since leave-one-out cross validation does not have the mechanism for random partition, the independent sample approach will be relatively simple. Note that the difference of the predictions of an instance by two classification algorithms will not be approximately normally distributed. Deriving the sampling distribution of the matched sample approach for leave-one-out cross validation is therefore not so straightforward.

In a data set containing n instances, let r_i be the number of correct predictions made by classification algorithm i evaluated by leave-one-cross validation for $i=A, B$. If $r_i \geq 5$ and $n-r_i \geq 5$ for $i=A, B$, then by the same argument as given in the proof of Theorem 3, $\overline{p}_A-\overline{p}_B$ will approximately follow a normal distribution with mean p_A-p_B and variance $(p_A(1\text{-}p_A)/n)+(p_B(1\text{-}p_B)/n)$ when the scheme-independence assumption is true for algorithms A and B. So, when the independent sample approach is used, k-fold and leave-one-out cross validation actually have the same sampling distribution for making statistical inference. Alternatively, when the matched sample approach is used, let x_{ij} represent the random variable corresponding to the prediction of the jth instance by classification algorithm i for $i=A, B$ and $j=1, 2, \ldots, n$. Then the value of $y_j=x_{Aj}-x_{Bj}$ can be -1, 0, or $+1$.

Theorem 4. *Let n_{-1}, n_0, and n_{+1} be the frequency of $y_j=-1$, 0, and $+1$, respectively. If $n_{-1} \geq 5$, $n_0 \geq 5$, and $n_{+1} \geq 5$, then the sampling distribution of $\overline{y}=\sum_{j=1}^{n} y_j/n$ can be approximated by a normal distribution with mean p_A-p_B when the scheme-independence assumption is satisfied.*

Proof. When $n_{-1} \geq 5$, $n_0 \geq 5$, and $n_{+1} \geq 5$, by the central limit theorem, point estimator $\overline{y}$ can be assumed to follow a normal distribution. Since $P\{x_{ij}=1\}=p_i$ and $P\{x_{ij}=0\}=1-p_i$ for $i=A, B$, and the scheme-independence assumption is satisfied, we have $P\{y_j=-1\}=(1-p_A)p_B$, $P\{y_j=0\}=(1-p_A)(1-p_B)+p_Ap_B$, and $P\{y_j=+1\}=p_A(1-p_B)$, and hence $E(y_j)=p_A-p_B$. Since every y_j is an unbiased estimator of p_A-p_B, $E(\overline{y})=\sum_{j=1}^{n} E(y_j)/n=p_A-p_B$.

Theorem 4 shows that the matched sample approach is applicable for leave-one-out cross validation. Sample variance $s_y^2=\sum_{j=1}^{n}(y_j-\overline{y})/(n-1)$ can be an estimate of $\mathrm{Var}(y_j)$ for $j=1, 2, \ldots, n$. The test statistic is therefore calculated as $t=\overline{y}/\sqrt{s_y^2/n}$ with $n-1$ degrees of freedom for testing null hypothesis H_0: $p_A-p_B=0$.

Example 2. Suppose that the number of correct predictions of leave-one-out cross validation for classification algorithms A and B on a data set with 100 instances is 80 and 84, respectively. In the independent sample approach, the test statistic for identifying whether the accuracies of the two classification algorithms are significantly different is calculated as

$$z=\frac{\overline{p}_A-\overline{p}_B}{\sqrt{2\overline{p}(1-\overline{p})/n}}=\frac{0.80-0.84}{\sqrt{2\times 0.82\times 0.18/100}}=-0.7362,$$

where $\overline{p}$ is the pooled mean accuracy. Let sample variance s_y^2 and the frequency of n_{-1}, n_0, and n_{+1} be 0.16, 30, 44, and 26, respectively in the matched sample approach. Then the test statistic for this case is

$$t=\frac{-0.04}{\sqrt{0.16/100}}=-1.0.$$

6. Conclusions

Both k-fold and leave-one-out cross validation are popular approaches for evaluating the performance of a classification algorithm, while how to use the testing results of the two evaluation approaches for making statistical inference are not all the same in literatures. In this paper, we consider four factors to investigate the usage of k-fold cross validation. The factors include the number of folds, the number of instances in a fold, the level of averaging, and the repetition of cross validation. In order to study the impact of the four factors, we first propose the independence assumptions and define the large-sample conditions in cross validation. They are then used to derive the sampling distributions of the point estimators for the two cross validation approaches in evaluating the performance of one algorithm or comparing the performance of two algorithms.

According to the sampling distributions for k-fold cross validation, the large-sample conditions determine the number of instances in a fold and the number of folds that can be as large as possible when the large-sample conditions still hold in every fold. If the variability of the performance of a classification algorithm is important, the level of averaging should be set at fold. Since the mean accuracies obtained by any two rounds of k-fold cross validation are dependent, repeatedly performing k-fold cross validation cannot provide a more reliable point estimate. Both the independent and the matched sample approaches can be used to make statistical inference about the testing results of leave-one-out cross validation.

When the scheme-independence assumption is not satisfied, neither the matched nor the independent sample approach can be applied to compare the performance of two classification algorithms for k-fold and leave-one-out cross validation. New statistical inference methods should be established to serve this purpose. Since there is a random partition mechanism in k-fold cross validation, how to reduce the variability of a point estimate obtained from k-fold cross validation is an interesting research topic. Repeating k-fold cross validation is not an appropriate way for this purpose, and hence new efficient methods should be developed to obtained more reliable accuracy estimates.

Conflict of interest statement

None declared.

Acknowledgments

This research was supported by the National Science Council Taiwan under Grant no. 101-2410-H-006-006.

References

[1] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann, Massachusetts, 2011.
[2] J.H. Friedman, On bias, variance, 0/1-loss, and the curse-of-dimensionality, Data Min. Knowl. Discov. 1 (1997) 55–77.
[3] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of K-fold cross-validation, J. Mach. Learn. Res. 5 (2004) 1089–1105.
[4] J.D. Rodriguez, A. Perez, J.A. Lozano, Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE. Trans. Pattern. Anal. Mach. Intell. 32 (2010) 569–575.
[5] E. Alpaydin, Introduction to Machine Learning, 2nd Edition, MIT Press, Massachusetts, 2010.
[6] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann, Massachusetts, 2012.
[7] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, 2nd Edition, John Wiley & Sons, New Jersey, 2011.
[8] T.M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

Please cite this article as: T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.03.009

[9] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison Wesley, Massachusetts, 2006.
[10] G. Casella, R.L. Berger, Statistical Inference, 2nd Edition, Duxbury, California, 2002.
[11] R.J. Freund, W.J. Wilson, D.L. Mohr, Statistical Methods, 3rd Edition, Academic Press, Massachusetts, 2010.
[12] D.R. Anderson, D.J. Sweeney, T.A. Williams, J.D. Camm, J.J. Cochran, Statistics for Business and Economics, 12th Edition, South-Western, Tennessee, 2012.

Tzu-Tsung Wong is a professor in the Institute of Information Management at National Cheng Kung University, Taiwan, ROC. He received his Ph.D. degree majored in industrial engineering from the University of Wisconsin at Madison. His research interests include Bayesian statistical analysis, naïve Bayesian classifiers, and classification methods for gene sequence data.

Please cite this article as: T.-T. Wong, Performance evaluation of classification algorithms by *k*-fold and leave-one-out cross validation, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.03.009