Improved Version of Monte Carlo 1 Algorithm

M. Hossin[1,3], M.N. Sulaiman[1], N. Mustapha[1], and R.W. Rahmat[2]

[1] Department of Computer Science, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 Serdang, Selangor Darul Ehsan, Malaysia. (mhossi78@gmail.com, nasir, norwati@fsktm.upm.edu.my)
[2] Department of Multimedia, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 Serdang, Selangor Darul Ehsan, Malaysia. (rahmita@fsktm.upm.edu.my)
[3] Department of Cognitive Science, Faculty of Cognitive Sciences and Human Development, University Malaysia Sarawak, 93400 Kota Samarahan, Sarawak, Malaysia. (hmohamma@fskpm.unimas.my)

I-Introduction

The nearest neighbor classifier (NNC) is one of the most popular and also the simplest data classification method. However, this method suffered from two main drawbacks which are large storage of prototypes and large response time classification process [1][3]. Due to these advantages, many instance selection (IS) methods have been proposed to improve and preserve the NNC performance competency by reducing the number of prototypes as much as possible and achieving as lowest as possible testing error Te_{err} simultaneously. The comprehensive discussion on specific IS methods can be found at [1], [2], [3], [5], [6], [7], and [8]. Among all, Monte Carlo 1 (MC1) [8] method is one of the simplest IS methods which have been designed to improve the performance of NNC. Although this method is simple and able to obtain good results, however, this method still has some limitations. Firstly, the selection of the best prototype sample is totally relied on prototype sample that has the lowest training error Tr_{err}. Unfortunately, this selected prototype sample does not always produce good generalization ability most of the time. Furthermore, the prototype searching also has been limited to 100 samples. Typically as a method that behaves stochastically and employs random selection this restricted n sample is not sufficient to search good prototype samples which can produce good generalization ability. Additionally, by using a smaller number of prototype samples this method is easy to be trapped at the local optimum. Therefore, this paper proposes a Modified Monte Carlo 1 method (MMC1) which proposed two modifications to increase the effectiveness of MC1 method. For more details the rest of the paper is organized as follows. The next section will discuss in brief about the proposed method. Meanwhile, the experimental results will be addressed in the following section. Finally, this paper ended with conclusion.

II-Material and Methods

In general, the original MC1 method proposed by Skalak [8] was based on repeated stochastic behavior of Monte Carlo method. As stated in [8], this method requires three fixed input parameters which are k is the value of k-NN, m is the number of prototype size (w^m) and n is the number of w^m that will be generated (w^{nm}). The implementation of this method is simply using repeated sampling of prototype samples by randomly pick the instances from a pool of $X = \{x^1, \dots, x^n\}, x^i \in \Re^p$. As used in [8], only $n = 100$ samples have been generated where each class has been represented by one prototype $W = \{w^1, \dots, w^c\}, w^i \subset \Re^p$ where c indicates the number of classes and $c > 1$. The best prototype sample W^* which has the lowest training error Tr_{err} will be employed for testing with $Z = \{z^1, \dots, z^n\}, z^i \in \Re^p$. More details about this method can be found in [8].

As stated earlier, this method only evaluates the best prototype sample W^* that has the lowest training error Tr_{err}. From the observation we made, sometime the best W^* with the lowest Tr_{err} might produce poor testing error Te_{err} if compare to the other prototype samples which have a little bit higher Tr_{err}. Even if two prototype samples have the same Tr_{err} , they can obtain different Te_{err} when dealing with Z. Clearly, this indicates that the single Tr_{err} measurement could not be the best measurement to determine the best W^*. Thus, we proposed three additional measurements to resolve the above problem. In this study, we proposed a modified version of precision P^c and recall R^c, and also mean squared error