Expert Systems with Applications 39 (2012) 11880–11888

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine

Chin Heng Wan [a,1], Lam Hong Lee [b,*], Rajprasad Rajkumar [b,2], Dino Isa [b,3]

[a] Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, 31900 Kampar, Perak, Malaysia
[b] Intelligent Systems Research Group, Faculty of Engineering, The University of Nottingham, Malaysia Campus, Jalan Broga, 43500 Semenyih, Selangor, Malaysia

ARTICLE INFO

Keywords:
Text document classification
K-nearest neighbor
Support vector machine
Euclidean distance function

ABSTRACT

This work implements a new text document classifier by integrating the K-nearest neighbor (KNN) classification approach with the support vector machine (SVM) training algorithm. The proposed Nearest Neighbor-Support Vector Machine hybrid classification approach is coined as SVM-NN. The KNN has been reported as one of the widely used text classification approaches due to its simplicity and efficiency in handling various types of text classification tasks. However, there exists a major problem of the KNN in determining the appropriate value for parameter K in order to guarantee high classification effectiveness. This is due to the fact that the selection of the value of parameter K has high impact on the accuracy of the KNN classifier. Other than determining the optimal value of parameter K, the KNN is also a lazy learning method which keeps the entire training samples until classification time. Hence, the computational process of the KNN has become intensive when the value of parameter K increases. In this paper, we propose the SVM-NN hybrid classification approach with the objective that to minimize the impact of parameter on classification accuracy. In the training stage, the SVM is utilized to reduce the training samples for each of the available categories to their support vectors (SVs). The SVs from different categories are used as the training data of nearest neighbor classification algorithm in which the Euclidean distance function is used to calculate the average distance between the testing data point to each set of SVs of different categories. The classification decision is made based on the category which has the shortest average distance between its SVs and the testing data point. The experiments on several benchmark text datasets show that the classification accuracy of the SVM-NN approach has low impact on the value of parameter, as compared to the conventional KNN classification model.

1. Introduction

Text document classification is defined as the task of assigning electronic text documents into relevant categories. In order to determine the appropriate category for an unlabeled text document, a classifier is used to perform the task of automatically classifying the text document. The rapid growth of the internet and computer technologies has caused the existence of billions of electronic text documents which are created, edited, and stored in digital ways. This situation has brought great challenge to the public, specifically the computer users in searching, organizing, and storing these documents. Therefore, text document classification approaches have gained their great importance as tools to search, organize, and store these huge amount of text data. In recent years, many text document classification approaches have been developed to encounter the problems in allocating text documents into their annotated categories. These approaches include K-nearest neighbor (Han, Karypis, & Kumar, 1999; He, Tan, & Tan, 2003), support vector machine (Bosnic & Kononenko, 2008; Diederich, Kindermann, Leopold, & Paass, 2003; Hao, Chiang, & Lin, 2009; He et al., 2003; Isa, Lee, Kallimani, & Rajkumar, 2008; Joachims, 1998, 1999, 2002; Lee, Wan, Rajkumar, & Isa, 2011a), decision tree induction (Greiner & Schaffer, 2001), maximum entropy (Nigam, Lafferty, & McCallum, 1999), rule induction (Apte, Damerau, & Weiss, 1994a, 1994b), Bayesian (Androutsopoulos, Koutsias, Chandrinos, & Spyropoulos, 2000; Chen, Huang, Tian, & Qu, 2009; Domingos & Pazzani, 1997; Eyheramendy, Genkin, Ju, Lewis, & Madigan, 2003; Kim, Rim, Yook, & Lim, 2002; Lee, Isa, Choo, & Chue, 2010a, 2011b; Lee & Isa, 2010; McCallum & Nigam, 1998; O'Brien & Vogel, 2003; Sahami, Dumais, Heckerman, & Horvitz, 1998), artificial neural networks (Chen, Lee, & Hwang, 2005; Soltanizadeh & Shariar, 2008), self-organizing maps (Isa, Kallimani, & Lee, 2009; Lee & Yang, 2003) and many other machine learning and statistical approaches.

* Corresponding author. Fax: +603 89248017.
E-mail addresses: wanchinheng@yahoo.com (C.H. Wan), leelamhong@gmail.com (L.H. Lee), Rajprasad.Rajkumar@nottingham.edu.my (R. Rajkumar), Dino.Isa@nottingham.edu.my (D. Isa).
[1] Tel.: +605 4688888; fax: +605 4661672.
[2] Tel.: +603 89248377; fax: +603 89248017.
[3] Tel.: +603 89248116; fax: +603 89248017.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.068

The K-nearest neighbor (KNN) classification approach has been widely used in various types of classification tasks (Han et al., 1999; He et al., 2003). This classification approach has gained its popularity based on its low implementation cost and high degree of classification effectiveness. However, the KNN has a unique requirement which has restricted its effectiveness and efficiency in handling classification tasks, which is the necessity in determining the appropriate value of parameter K in order to optimize the classification accuracy. This problem has been addressed by many research groups since the past decades (Geng et al., 2008). According to Geng et al., if the value of K is generally small, the performance of the KNN is moderate. This is due to the fact that the smaller the value of K, the lesser information are obtained from the training data and vice versa. The accuracy of the KNN classifier steadily increases if the value of K increases. However, after the value of k increases until a certain threshold, the performance of the KNN classifier decreases due to the fact that there are too many neighbors have been used, hence leads to the occurrence of noises to the information obtained from the training data (Callut, Saerens, & Dupont, 2008). This experimental result is illustrated in Fig. 1.

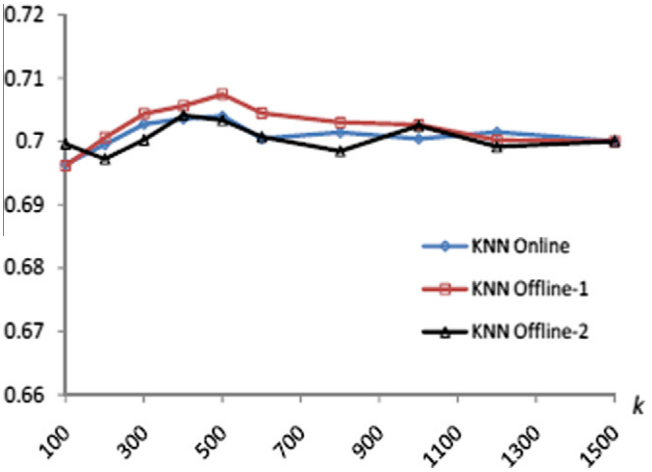

Fig. 1. Number of K will affect the performance of the KNN (Geng et al., 2008).

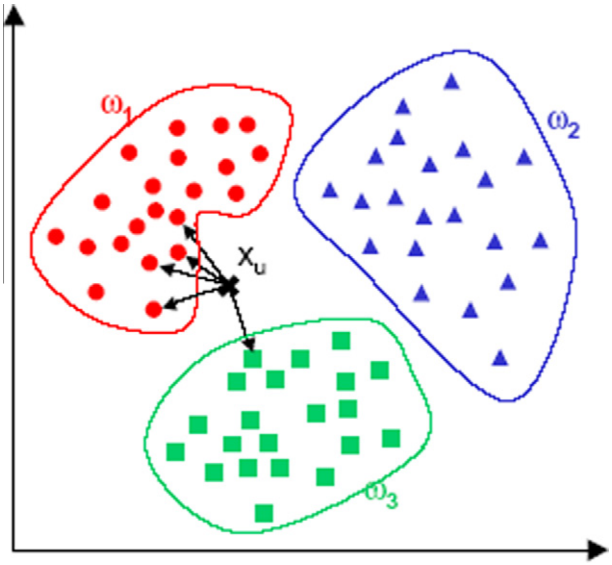

Fig. 2. Feature space of a 3-dimensional 5-NN classifier. (Source: Osuna, 2002).

In this paper, we present a hybrid classification approach by incorporating the support vector machine (SVM) training algorithm to the training stage of the KNN approach, coined as SVM-NN, in order to overcome the problem of the necessity in determining the optimal value for parameter. In the SVM-NN hybrid classification model, the SVM is used to reduce the training samples of the classifier to the support vectors (SVs) of each category, and the nearest neighbor algorithm is then used to compute the average distance between the testing data point to the set of SVs from different categories. The classification decision is made based on the category which has the shortest average distance between its SVs with the testing data point. As the proposed SVM-NN approach computes the distance from the testing data point to all the data points which have been identified as the SVs for each different category, the determination of optimal value of parameter K in the conventional KNN classifier could be ignored. The experimental results show that the accuracy of the SVM-NN classifier has low impact on the implementation of parameter, and yet the performance is comparable to the conventional KNN classifier.

2. K-nearest neighbor classification approach

The K-nearest neighbor (KNN) classification approach is an instant-based learning algorithm that uses the nearest distance in determining the category of new vector in the training data set (Han et al., 1999). During the training stage, the feature space is divided into multiple regions and the training data points are mapped into these regions according to the similarity of their contents. The unlabeled input data points are categorized to a particular category by finding the closet or distance from input data point and that particular category. The KNN approach needs only a small number of training data points and this has contributed to the simplicity of the KNN which makes it outperforms other classification approaches (Osuna, 2002). Fig. 2 shows an example of a 5-NN classifier which consists of three categories ω_1, ω_2, and ω_3. X_u is the new unlabeled input data point to be classified in the testing stage.

The most commonly and widely used distance function for the KNN classifier is the Euclidean distance formula and it is used to calculate the distance between the new unlabeled data point and the training data points. The main step in the classification stage of the KNN is to measure the distance in order to identify the nearest neighbors of the new input data point (Han et al., 1999).

According to Fig. 2, the value of parameter K is 5 and Euclidean distance formula has been used to calculate the distance between the training data points and the testing data point $X_{u.}$ Among the five nearest neighbors of X_u, four are belong to category ω_1 and another one belongs to category ω_3. Hence, X_u is classified as category ω_1 by the KNN classifier (Osuna, 2002). The main advantage of the KNN is that it can perform well in classification tasks with multi-categorized data points. On the other hand, since the KNN uses the distance calculation in determining the category of new input data points, this has brought to a great disadvantage of the KNN when the size of training set is big, where the classification model will become computationally intensive. This problem has made the KNN demands for more memory and CPU usages and high time consumption, especially in the classification stage. Moreover, this problem has also led to the drastic decrease in accuracy when there exist irrelevant features or noises in the training dataset.

As an instant-based learning approach, the KNN freezes the training process until it receives a new input data point to be classified. The KNN compiles the entire training data points again when there is a new input sample and it discards the immediate result (Osuna, 2002). To counter the major drawback of the KNN approach, which is high time consumption, several enhancement techniques such as bucketing algorithm (Osuna, 2002) and k-dimensional trees algorithm (Osuna, 2002) have been introduced to improve the