**4 uncertainty and evidence theory based kNN algorithm (UCEkNN)**

In this section, we present our own kNN algorithm based on uncertainty and evidence theory(UCEkNN), and then we proposed 2 optimal k value selection algorithm based on the UCEkNN.

**4.1 UCEkNN**

According to equation (5), till now we use to represent the BPA of one neighbor xi among k nearest neighbors set whose class is Cq. In practice, however, there is some uncertainty of the label itself. While the BPA of each neighbor is combined, the uncertainties are also accumulated, and the accuracy of the ultimate prediction will decrease to some extent. From this protective, the uncertainty of the label can not be ignored. we define the uncertainty as UC:

**Definition 2: UCi whose value ranges from 0 to 1 describes the uncertainty of the label itself of the neighbor xi among the k nearest neighbors of xs . UCi is depended and only depended by the neighbor's characteristic.**

According to definition 2 and equation 5, we proposed a new form for the BPA function:

(9)

Similarly, according to DS combination rule, for each class q:

(10)

Although UCi has been discussed in [3] as the imperfect labeling, our own approach is different intrinsically. In our approach, we do not change the basic form of equation (5), i.e. we assume one neighbor only belongs to one specific class in spite of uncertainty rather than two or more possible classes. Moreover, we think UCi is independent with the distance, and can be accessed easily with some extra information of xi. Here we use a opinion set to describe the extra information, we define it as OpS:

**Definition 3: OpSi describes the extra information set of xi, it can be a set of votes information, ratings information, or other reasonable information which illustrate experts' or users' subjective opinions on wether xi should be labeled with Cq.**

In practical applications, for the kNN algorithm, the label of each sample is decided by a group of people, usually by authoritative experts. OpSi is such a set which contains the opinions. OpS is accessible in many problems. For example, [9] introduce a medical therapy recommendation system based on kNN algorithm, for each patient in the training data set, the ultimate therapy is voted by at least 5 doctors. The most voted therapy is used to label this patient. Here, the vote results of doctors' make up the OpS of this patient.

Intuitively, the consistency of doctors' opinions reflects the complexity of making decisions. If all doctors vote to the same therapy, the case of that patient is not complex and clear to make a decision, while for a multi-vote-result situation, the case of that patient can be sophisticated and controversial so that the divergence come out.

We use the information entropy(IE) [12] of OpSi to quantify its consistency:

(11)

where Pic is the proportion of class c in OpSi.

Moreover, the uncertainty UCi of the label of xi has some connection with the uncertainty of OpSi. For a high-consistency case where H(OpSi) has a small value, the UCi should be more close to 1 while for a controversial case where H(OpSi) has a large value, he UCi should be more close to 0. From this prospective, we conclude the UCi is a decreasing function of H(OpSi). We suggest to chose the following function:

(12)

For a query instance of unknown category xs, the predicting process is shown in Algorithm 1.

| Algorithm 1 Outline of the proposed Algorithm |
| --- |
| Input : a training set , a reasonable k, a query instance xs, OpSi for each sample in training set  Output : the optimal label of xs |
| 1 get the k nearest neighbors set  2 for every xi in T:  calculate H(OpSi) for each xi using equation (11)  calculate BPA for each xi using equation (9)  3 for each classification Cq in C:  calculate combining BPA ms({Cq}) of Cq  4 return the optimal label whose BPA ls the largest |

**4.2 optimal k value selection based on UCEkNN**

Traditional k value selection algorithm aim to find the global optimal k for a fixed training data set and testing dataset.Although a good value might be obtained using cross validation (CV), the same value is unlikely to be optimal for the whole space spanned by the training set. In this brief, we devised a new greedy method based on UCEkNN.

For different k ranging in [kmin,kmax], the neighbors set of xs is different, so is the BPA of different classification. Consequently, to search a optimal k for a unlabeled sample is tantamount to determine a best opportunity when the BPA of all classifications meet some optimal condition. Here, we give two available conditions:

**condition 1:** for different k, the largest value of ms({Cq}) reaches the maximum

**condition 2:** for different k, the difference between the largest and second largest value of ms({Cq}) reaches the maximum

Detailed execution steps of our approach under condition 1 and condition 2 are illustrated by Algorithm 2 and Algorithm 3 respectively.

| Algorithm 2 Outline of the proposed Algorithm |
| --- |
| Input : a training set , a query instance xs, OpSi for each sample in training set, a minimum value of k, kmin, and a maximum value of k, kmax  Output : the optimal k |
| 1 for each k in [kmin,kmax]:  2 for every xi in T:  calculate H(OpSi) for each xi using equation (11)  calculate BPA for each xi using equation (9)  3 for each classification Cq in C:  calculate combining BPA ms({Cq}) of Cq  4 get  5 return the k value whose ms(A) is the largest |

| Algorithm 3 Outline of the proposed Algorithm |
| --- |
| Input : a training set , a query instance xs, OpSi for each sample in training set, a minimum value of k, kmin, and a maximum value of k, kmax  Output : the optimal k |
| 1 for each k in [kmin,kmax]:  2 for every xi in T:  calculate H(OpSi) for each xi using equation (11)  calculate BPA for each xi using equation (9)  3 for each classification Cq in C:  calculate combining BPA ms({Cq}) of Cq  4 get  5 get  5 return the k value whose ms(A)-ms(B) is the largest |

**4.3 optimization of UCEkNN using L-Sure algorithm**

Actually, the influence of the opinion about the data sample's label from different experts or users should not be identical. For instance, in a medical decision, the opinion of an experienced and proficient doctor could be more crucial. From this prospective, just using entropy of OpS to measure the uncertainty is not reasonable any more.

L-Sure algorithm is devised to intensify the opinion from more authoritative people who vote for the label of a sample. The basic idea of L-Sure algorithm can be described as two steps: (1) select L most authoritative experts out of the vote set. (2) calculate the uncertainty based on the opinion of these L experts.

Given an unlabeled sample and its k neighbors, a straightforward strategy to pick the L most reliable experts is just sort all the experts by the precision in the k neighbor from high to low. Such a idea is easy but efficient.

After getting the L experts for an unlabeled sample, the uncertainty can be calculated by the following condition: (1) if all the L experts hold the same opinion, then the UCi in equation (9) equals 1 (2) if one or more experts out of L experts hold different opinions, then we still use eqution (12) to calculate the UCi.

All the above steps can be described as equation (13):