# K-Nearest Neighbor for Uncertain Data

Rashmi Agrawal
Research Scholar, ManavRachna International University, Faridabad

## ABSTRACT

The classifications of uncertain data become one of the tedious processes in the data-mining domain. The uncertain data are contains tuples with different data and thus to find similar class of tuples is a complex process. The attributes which have a higher level of uncertainty needs to be treated differently as compared to the attributes having lower level of uncertainty. Different algorithms exist in literature for users to choose a suitable one as per their need. This research paper deals with the fundamentals of various existing data classification techniques for uncertain data using the k nearest neighbor approach. The literature shows that much work has been done in this area but still there are certain performance issues in the k nearest neighbor classifier. K nearest neighbor is one of the important algorithms in top 10 data mining algorithms.

## Keywords

Data Mining, Classification, Uncertain Data, Nearest Neighbor, Probability

## 1. INTRODUCTION

The commercial and research interests in data mining is increasing rapidly, as the amount of data generated and stored in databases of organizations is already enormous and continuing to grow very fast [1]. As more data are gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. Data Mining refers to the non-trivial extraction of implicit, previously unknown and potentially useful information from data in databases. Classification is a well-recognized Data Mining task and it has been studied extensively in the fields of statistics, pattern recognition, decision theory, machine learning, neural networks and more. Classification operation usually uses supervised learning methods that induce a classification model from a database. The task of classification is to assign a new object to a class from a given set of classes based on the attribute values of the object [2]. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects [3, 9]. For data classification, the well-known algorithm of top-10 data mining algorithm is K-NN classifier. The *k*-Nearest Neighbor (*k*-NN) is one of the simplest classification methods used in data mining and machine learning.

## 2. CLASSIFICATION

Classification is one of the important techniques in data mining which is used to classify the data into the specified classes. In this technique, a training set is used where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model, used to classify new objects. Various techniques like decision tree, K nearest neighbor, Naïve Bayes method , Neural network are used to classify new objects.

A decision tree is a simple recursive structure for expressing a sequential classification process in which a case described by a set of attributes is assigned to one of a disjoint set of classes. Various efficient algorithms have been developed to construct a decision tree in a reasonable amount of time. Generally, these algorithms (ID3, C4.5, CART, SPRINT) follow the greedy approach by taking the local optimum decision. Usually information gain, entropy, gini index and other similar measures are used to take the decision of selecting the attribute in split point.

A neural network is a computational model based on the biological neurons. It is abstracted as a directed graph where the neurons represent the nodes and connections between them are edges. The weight on each edge represents the inhibiting or stimulating type and the strength of interaction between the neurons. The neural network is trained to respond with certain inputs to produce the desired output. Back propagation network and other neural network architectures are used in machine learning and pattern recognition applications.

A Naïve Bayes classifier is a simple probabilistic classifier and works well for many applications specially in text classification. Naïve bayes is a supervised and statistical learning method. This classifier is based on the bayes theorem with naïve independence assumptions. A naïve bayes classifier assumes the presence or absence of a feature of a class is not related to any other feature of the class. This method is popular for many reasons. It is very easy to construct and does not need any complicated iterative parameter estimation schemes. Also it may be applied to large data sets. It is easy to interpret and sometimes the outcomes of this classifier are surprisingly well.

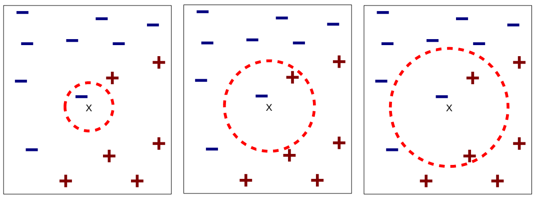
A rule-based classifier is a technique for classifying records using a collection of ”if ... then ...” rules. a rule-based classier consists of a set of rules, used in a given order during the predictionprocess, to classify unlabeled objects. A rule r covers an instance x if the attributes of the instance satisfy the condition (LHS) of the rule. Quality of a classification rule can be evaluated by two parameters coverage and accuracy. Coverage is the fraction of records that satisfy the antecedent of a rule. Accuracy can be defined as the fraction of records covered by the rule that belong to the class of right hand side.

The K-Nearest Neighbor (K-NN) is one of the simplest classification methods used in data mining. It makes the classification by getting votes of the k Nearest Neighbors. In this paper our focus will be on KNN classifier. In the next section we will study the KNN classifier.

## 3. KNN ALGORITHM

K- Nearest Neighbor (KNN) classifier comes in the category of a lazy learner. A lazy learner stores the given training tuple and does nothing and waits for a test tuple. In the KNN classification approach all training tuples are stored in an n-dimensional space. When a tuple from the test data is given to the classifier, it searches the k training tuples which are closest to the unknown tuple. These selected k tuples are k nearest neighbor of the unknown tuple.

To classify an unknown record the distance between other training records are computed. Based on the distance the K nearest neighbors are identified and class labels of these nearest neighbors is used to determine the class label of unknown record. The following figure represents the example of K nearest neighbor.



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

**Figure 1: K nearest neighbor**

The closeness or nearness of a tuple is defined by the distance metric. One commonly used distance metric is Euclidean distance and is calculated as

$$d(p,q) = \sqrt{} \sum (pi - qi)2$$

The other commonly used distance metric is Manhattan distance. For the categorical variable hamming distance is used.

The basic KNN method is unable to handle the uncertainty and imprecision in the labeling of known classes. This can lead to a problem as in real life in many cases uncertainty arises. In engineering applications like location based services, biological management system, remote sensing the records in the database are generally uncertain. For example to diagnose correctly a patient a doctor may have uncertain values in the blood or stool samples. In such databases basic KNN method is not applicable. To cope with this issue various researchers have proposed the modified KNN approach which can handle uncertain data. In the following section we will discuss some of the major contributions in the classification of uncertain data using the KNN method.

## 4. KNN APPROACH FOR UNCERTAIN DATA

Liu and pan[4] proposed a fuzzy belief K- nearest neighbor (FBKNN) method to handle uncertain data. It uses the concept of meta classes to characterize the classes which are difficult to classify in uncertain data. Fuzzy membership is assigned to each labeled training sample. Based on this a new input sample is classified using the KNN. Using the assigned memberships of the KNNs and appropriate distance measure K basic belief assignments are created and then these K basic belief assignments were fused. They evaluated the method on four real data sets and performed k fold cross validation by different classification methods. They proved that FBKNN method is convenient and efficient for engineering applications.

Fabrizio and Fassetti [5] argued that the approaches which uses the distance measures like distance between means, the expected distance, and probabilistic threshold distance is a naïve approach and is not capable of producing the exact result and may predict the wrong class even if the probability of belongingness of an object to a certain class is zero. They proposed the uncertain nearest neighbor (UNN) rule. This rule depends on the nearest class rather than nearest object. In the K- nearest neighbor rule whenever an object is assigned to a class it is checked by finding the maximum members present among its K-nearest neighbors in the training set. Their contributions in the proposed method can be summarized as follows-

1) They proposed the concept of nearest neighbor class.
2) It was shown that nearest neighbor class concept is more powerful than nearest neighbor object.
3) Based on this uncertain nearest neighbor classification rule(UNN) was defined.
4) An algorithm was designed to implement this.
5) By the experimental results it was proved that UNN rule was able to classify in more effective way.

Jianping Gou *et al*[6]. have presented a local mean-based k-nearest centroid neighbor classifier that assigns to each query pattern a class label with nearest local centroid mean vector so as to improve the classification performance. The presented scheme not only takes into account the proximity and spatial distribution of k neighbors, but also utilized the local mean vector of k neighbors from each class in making classification decision. In the presented classifier, a local mean vector of k nearest centroid neighbors from each class for a query pattern was well positioned to sufficiently capture the class distribution information. In order to investigate the classification behavior of the presnted classifier, they conducted extensive experiments on the real and synthetic data sets in terms of the classification error. Experimental results demonstrated that their presented method performed significantly well, particularly in the small sample size cases, compared with the state-of-the-art KNN-based algorithms.

Destercke [7] in 2012 proposed an approach for k nearest neighbor classification that can handle uncertain data in a very generic way. They considered the coherent lower previsions and find different neighbors as information sources that provide non-totally reliable information modeled by coherent lower previsions. They argued that a lower prevision approach with an imprecise decision rule naturally tackles the problem of distant neighbors.

Chang and Chen in 2009[8] proposed the probabilistic threshold k nearest neighbor query (T-K-PNN). They argued that a user may require answers with confidence that meets some threshold condition. They presented three methods to process T-K-PNN. In the first method which was named as k bound filtering all objects which were not having a chance to be selected in a query were removed. The second method was proposed as probabilistic candidate selection by finding a smaller set of candidates. In the third method two kinds of verification were done lower bound based and upper bound based. They argued that the solution reduces input output overhead by using R-Tree which can prune a large number of objects.

Agarwal, Arnor and Phillips studied in 2013[10] the nearest neighbor queries in a probabilistic framework. The location of each input point was specified as a probability distrinution function. They guaranteed the non-zero NNs and presented two algorithms for computing qualification probabilities efficiently. A structure called the probabilistic voronoi diagram was build and a simple Monte Carlo approach was used to build an index for quick computation. It was shownthrough the experiments that the Monte Carlo method and spiral search methods were highly effective in computing the nearest neighbors.

K-NN classifier may have problem incase training samples are uneven. The difficulty with KNN classifier is that it decrease the precision of classification in case of uneven density of training data.Lijuan Zhou [11] proposed clustering-based K-NN method. It predefine training data with the help of clustering method, and then they classify with a new KNN algorithm which implement a dynamic adjustment in each iteration for neighborhood number K. This proposed method helps to avoid uneven classification phenomenon and helps to decrease the miscalculation of boundary testing samples.

In case if there are an infinite number of samples in training set, then the possible outcome from the nearest neighbor classification (kNN) is independent on its assumed distance metric. But it is unfeasible that that the number of training samples is infinite. Hence, selecting distance metric becomes major issue in deciding the performance of KNN. YunlongGaoa [14] proposed two-level nearest neighbor algorithm (TLNN) in order to decrease the mean-absolute error of the misclassification rate of kNN with finite and infinite number of training samples. At low-level, Euclidean distance is used in order to determine a local subspace centered at an unlabeled test sample. At high level, they used AdaBoost as guidance to extract local information. Here data variance was maintained with TLNN method and also highly stretched or elongated neighborhoods were generated along different direction. The TLNN decrease the extreme dependence on the statistical method, which realized former knowledge from training data. Also the linear combination of few base classifier generated by weak learner in AdaBoost can produce much better kNN classifiers.

In the k-NN algorithm, the predefined k values neglect the influence of category and document number of training text. Hence, choosing the accurate value of K can attain better classification results.An Gong and Yanan Liu [13] proposed a type of dynamic attain k-valued for kNN classification algorithm, their experimental results proves that the dynamic attain k-valued kNN classification algorithm with high performance. While dealing with excessive data, a significant disadvantage of existing kNN algorithm is that the class with more frequent samples tends to govern the neighborhood of test request irrespective of distance measurements, which results in suboptimal classification performance on minority class. In order to solve this problem, Wei Liu,WeiLiu,SanjayChawla [12] proposed CCW (class confidence weight) which make use probability of attribute value updated in the class labels to weight prototypes in k-NN. The main benefit of using CCW is that it is able to precise the inherent bias to majority class in existing k-NN algorithm on any distance measurement. This is proved by theoretical analysis and comprehensive experiments.

## 5. ISSUES IN KNN

There are various performance issues that can affect the performance of k nearest neighbor classifier.

a. The first and most important issue is selection of k. if it is too large, it may select too many irrelevant data object as nearest neighbors. On the other hand if k is too small then it may not provide the correct result.

b. Another important issue is choice of distance measures. Some distance measures can be affected by the dimensionality of the data.

c. Finally the KNN classifiers are lazy learners therefore it is relatively expensive to classify unknown objects using the k nearest neighbor approach.

## 6. CONCLUSION

In terms of performance the algorithms developed so far for precise data in different data mining techniques like classification, clustering and association rule mining, we get satisfactory results but the uncertain data provides a completely different scenario and most algorithms give different results when applied on this data. In this paper we have studied few techniques of K nearest neighbor classification algorithms on uncertain data. Uncertain data mining is an area of interest for researchers and much work is required to handle the uncertain data in a better way. We further plan to go into the detail of classification technique for uncertain data to produce the accurate results as possible with the certain data.

## 7. REFERENCES

[1] Zhou Y., Youwen L., Shixiong X., An Improved KNN Text Classification Algorithm Based on Clustering, *Journal of Computers*, 2009, 4(3): 230-237.

[2] Romero C., Ventura S., Espejo P.G., and Hervas C., Data Mining Algorithms to Classify Students, Proceedings of the 1st Int'l conference on educational data mining, Canada, 2008, pp: 8-17.

[3] Zhang J., Mani I., kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction, In Proceedings of The Twentieth International Conference on Machine Learning (ICML-2003), Workshop on Learning from Imbalanced Data Sets II, August 21, 2003.

[4] Z.G. Liu, Q. Pan, J. Dezert. “A new belief-based K-nearest neighbor classification method,” Pattern.Recogn., vol. 46, No. 3, pp. 834-844, March, 2013

[5] FabrizioAngiulli, Fabio Fassetti,” Nearest Neighbor-Based Classification of Uncertain Data,” Journal ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 7,No.1, March 2013.

[6] JianpingGou,ZhangYi,Lan Du andTaisongXiong,"A Local Mean-Based k-Nearest Centroid Neighbor Classifier," The computer journal, Vol.54,No.1,January 2012.

[7] Destercke S, A k-nearest neighbours method based on imprecise probabilities. Soft Comput 16(5):833–844, 2012

[8] ReynoldCheng , Lei Chen , Jinchuan Chen , XikeXie, Evaluating probability threshold k-nearest-neighbor queries over uncertain data, Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, March 24-26, 2009, Saint Petersburg, Russia

[9] Tan, Songbo. "An effective refinement strategy for KNN text classifier." Expert Systems with Applications 30.2 (2006): 290-298.

[10] Pankaj K. Agarwal, AlonEfrat, SwaminathanSankararaman, and WuzhouZhang.Nearest-neighbor searching under uncertainty.In PODS, 2012.

[11] Lijuan Zhou, Linshuang Wang, XuebinGe, Qian Shi, "A clustering-Based KNN improved algorithm CLKNN for text classification", 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR), vol. 3, pp. 212 - 215, 6-7 March 2010.

[12] Wei Liu, Sanjay Chawla, "Class Confidence Weighted KNN Algorithms for Imbalanced Data Sets",Advances in Knowledge Discovery and Data Mining Lecture Notes in Computer Science, Vol.6635, pp 345-356, 2011.

[13] An Gong, Yanan Liu, "Improved KNN Classification Algorithm by Dynamic Obtaining K", Advanced Research on Electronic Commerce, Web Application, and Communication, Communications in Computer and Information Science Volume 143, 2011, pp 320-324.

[14] YunlongGaoa, JinyanPanb, GuoliJia, ZijiangYangc, "A novel two-level nearest neighbor classification algorithm using an adaptive distance metric", Knowledge-Based Systems, Vol. 26, PP. 103–110, February 2012.