COMPARISON OF CRISP AND FUZZY KNN CLASSIFICATION ALGORITHMS

Faraj A. El-Mouadib[*] and Amal F. Abdalsaalam[**]
[*]University of Garyounis, Faculty of Information Technology, Benghazi, Libya
elmouadib@yahoo.com ,

[**]Bank of Commerce and Development, Benghazi, Libya
Amal22032008@yahoo.com

ABSTRACT

Recently, with the advances in information technology tools in the collection and generation of massive amounts of data had overwhelmed the human ability and traditional data analysis capabilities to analyze and use such data. The cognitive science field, learning theory and human learning processes are essential areas to create new intelligent computer systems that seek to execute tasks similar to natural human tasks such as classification. Such needs had called for the development of new field such as Machine Learning and Knowledge Discovery in Databases to utilize the use and benefits from such data in the form of knowledge. Classification is one of the most important and well known tasks in the field of Data Mining.

In this paper, we focus on the Instant-Based Learning (IBL) classification method especially on the study of the K-nearest Neighbor classification algorithm from the crisp point view as well as the fuzzy. Computer system software is developed for the crisp and fuzzy K-nearest Neighbor classification algorithms with the introduction of the concept of Windowing of ID3. Our system is developed in Visual-Basic.net programming language. Some experiments as conducted by the use of well known data sets to conduct some comparison of the results.

Keywords: *Classification, Data mining, K-Nearest Neighbor classification algorithms, Cognitive Learning, Supervised learning, Unsupervised learning.*

1. INTRODUCTION

Humans as well as animals have superior pattern-recognition capabilities for tasks such as identifying faces, voices, smells, etc… So, many of human capabilities are gained through one of the learning methods that are available within the environment that they live in. Loosely speaking, learning is a process by which knowledge acquired through study, experience, or observation. According to [2, 18], the field of cognitive science increases our ability to understand how human can learn. The principles of cognitive science and different learning theories are implemented as new computer technologies to simulate the process of learning in humans and animals.

Nowadays, with the advances in information technology tools, massive amount of data have been collected and generated. These huge data have called for the development of new field known as Machine Learning (ML), in order to discover and learn useful knowledge from it. As stated in [18], the learning process can be seen as an information processing task that includes; the interpretation of sensory events, categorization of information, search of memory for past experiences and/or ideas, manipulation of ideas, images, and concepts. According to [1, 13], learning methods are categorized as; supervised and unsupervised. Supervised learning means the learning by the help of a supervisor or a teacher while unsupervised learning means learning without any help from external source.

With the great advances in computer technology, maintaining large volumes of data stored in databases became very easy and cheap. From these different sorts of data useful Knowledge can be discovered in many different forms such as; patterns, trends, regularities, anomalies, etc…

The stunning progress in the field of Artificial Intelligence (AI) is the Knowledge Discovery in Databases (KDD), which has evolved to include concepts from other fields such as; databases, machine learning, pattern recognition, statistics, information theory, reasoning with uncertainties, knowledge acquisition for expert systems, data visualization, machine discovery, and high performance computing. KDD is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. [6].

Data Mining (DM) is the most essential component of the process of KDD. In general and According to [8, 9 and 12], DM functionalities are categorized into predictive and descriptive. Predictive DM tasks are processes to classify or predict the class of un-classified objects (i.e. classification, prediction and estimation). Descriptive DM tasks are processes to discover hidden patterns or relationships in the data (i.e. characterization, clustering, discrimination, outlier analysis and association).

In this paper, we develop a software system to study in a comparison fashion the crisp and fuzzy K-Nearest Neighbor classification algorithms with some modification by adding the concept of windowing used in the ID3 algorithm. The system is developed in Visual basic.net programming language and it is tested with a number of well-known data sets in the form of experiments [14]. The comparison focuses mainly on the performance aspects of the crisp verses the fuzzy K-Nearest Neighbor classification.

2. CLASSIFICATION REVIEW

Classification is one of the old and very well known supervised learning tasks. According to [8], Classification is the processing of finding models or functions that describe and distinguish data classes or concepts. The deduced model or function is used to predict the class of objects whose class label is unknown. The derived models or function is based on the analysis of a set of training data (data objects whose class label is known). Classification is one of the most important human activities that help us to understand and communicate with our environment. Humans are always thinking about classifying all things surrounding them to be handled very easily and accurately.

Classification is an essential human activity; it is the corner stone of the cognitive science and the learning theory where humans learn to classify objects in some categories. The classification problem is to build an intelligent method to classify things in their prospective places. Such method is called a classifier that will be used to classify unclassified objects.

Generally speaking, any classification method is used to produce a classifier. An architecture of a typical classifier is depicts in figure-1. According to [3, 13, 15, and 20], the steps to construct a typical classifier are as follows:

1. The classification algorithm is provided with a set of pre-classified data objects (training set) that is used to construct the provisional classifier model.
2. A test set is used to examine the provisional model. The provisional model is adjusted to minimize the error rate (miss-classified cases).
3. The validation data set is used to readjust provisional model to produce the final classifier.

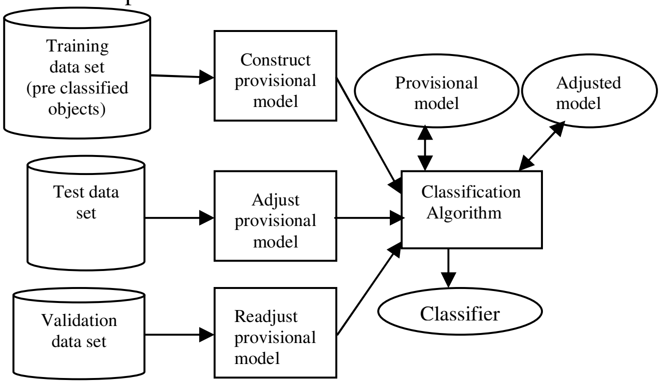

Figure-1: Architecture of a typical classifier.

In general, classification methods can be categorized as crisp or fuzzy. As it is stated in [4, 8, 10, 15, 17 and 19], a crisp classification method assigns each data object to one and only one class (full membership) based on the concepts of crisp logic. The *Membership function* (*Mf*) is defined by:

$Mf(x_i)=1$ or $Mf(x_i)=0$
If ($Mf(x_i)=1$) then the object x is in class i.
If ($Mf(x_i)=0$) then the object x is not in class i.

On the other hand, in fuzzy classification, each data object belongs to all of the classes with different degrees of membership. The total of all degrees of memberships for an object is equal to 1. The idea of fuzzy classification is based on the concepts of fuzzy logic. The *Membership function* (*Mf*) is given by:

[$0 \leq Mfx_i \leq 1$] where Mfx_i is the membership degree of object x in class i.

$$\sum_{i=1}^{c} Mfx_i = 1$$

is the total membership of object x in all of the C classes is equal to 1.

Classification has been around in use since the 1940s. There have been many methods and algorithms developed for classification with different methodologies and techniques. According to [4, 5, 7, 8, 11 and 14], there are many method and techniques for classification such as; Decision trees, Neural networks, Bayesian classifiers and Instance-Based Learners (IBL) methods.

IBL methods are based on the determination of the minimum distance between the new unclassified object and all of the objects in the training set. The class of unclassified object would then be the class of the nearest objects (The closest neighbor). The distance is computed by the Euclidean distance function or any other distance function.

3. K-NEAREST NEIGHBOR CLASSIFICATION ALGORITHMS

One of the earliest IBL algorithms (since the early 1950s) is the K-NN algorithm. According to [4, 5, 7, 12, 10 and 19], K-NN algorithm can be categorized as Crisp K-Nearest Neighbor (CK-NN) algorithm or Fuzzy K-Nearest Neighbor (FK-NN) algorithm. The similarity between the CK-NN and FK-NN algorithms is that both of them are used to assigning a class label to a newly unclassified data object. In the CK-NN algorithm, each newly unclassified object is assigned to the closest class with a full membership degree of 1. While the FK-NN algorithm is more suitable to classify ambiguous or uncertain data objects in the sense that each object belongs to all classes with different degrees membership.

4. DESIGN AND IMPLEMENTATION

Here we demonstrate the design and implementation of our system named CFK-NN, which consists of CK-NN and FK-NN subsystems. The CK-NN subsystem is a classifier based on Crisp K-Nearest Neighbor classification algorithm. The FK-NN subsystem is a classifier based on Fuzzy K-Nearest Neighbor classification algorithm.

4.1 ANALYSIS OF CFK-NN SYSTEM

The objects to be classified are represented in a table like format. The table is organized in rows and columns. Each row represents one object and each object is described by a number of attributes. For the training data set there is one additional attribute called the class attribute (class label). The aim of our classification system is to assign a class label for the new unclassified object (i.e. predict class label) by the use of the training data set. In general, the goal of our system is to build classifiers that can be used for prediction. Our system needs some special requirements that must be provided by the user so it can perform its functions accordingly such as:

- The name of the database that contains the pre-classified examples.
- The K value that represents the number nearest neighboring objects.
- The attributes to be used in the classification process.

The CFK-NN system is built in two phases. The first is to build the first classifier that is the CK-NN classifier. The final stage of CK-NN classifier is assumed to be the first step to build the second classifier that is the FK-NN classifier. After building the two classifiers then our system can be used to classify new unclassified objects.

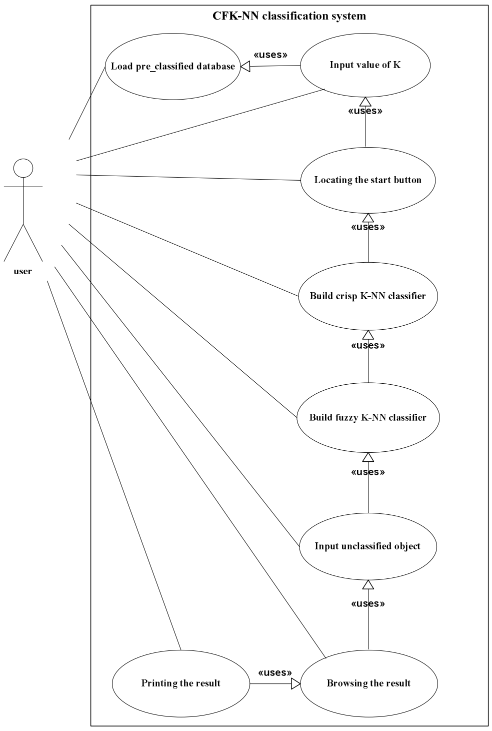
CFK-NN classification system
Load pre_classified database
«uses»
Input value of K
«uses»
Locating the start button
«uses»
Build crisp K-NN classifier
«uses»
Build fuzzy K-NN classifier
«uses»
Input unclassified object
«uses»
user
Printing the result
«uses»
Browsing the result

Figure-2: Use case diagram for CFK-NN classification system.

4.2 DESIGN OF THE CFK-NN CLASSIFICATION SYSTEM

We have We have used the Unified Modeling Language (UML) [7, 16], in the analysis and design phases. The Unified Modeling Language (UML) is a general modeling language consists of a set of diagrams that are used to describe, visualize, construct and document any software system. We have used the Visual Basic.NET 2005 programming language in the actual implementation of the system. Here, we will demonstrate only the use case diagram, activity diagram, class diagram and the Sequence diagram of our system, due to their importance and capabilities to give a clear view of the system.

The Use case diagram that describes the different functions of the system is depicted in figure-2.

Figure-3 and figure-4 depict the activity diagram for the CK-NN classifier and the activity diagram for FK-NN classifier respectively.

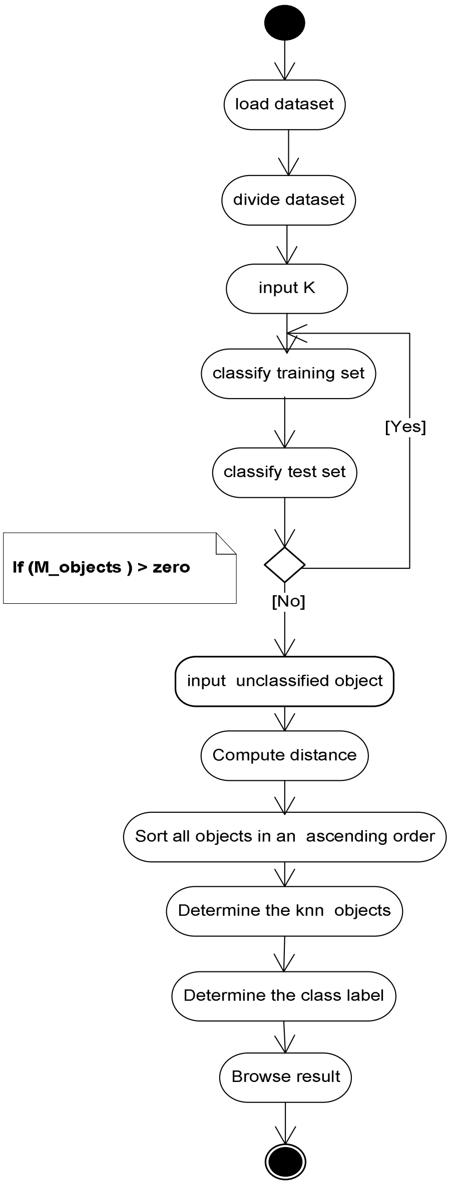

Figure-3: Activity diagram of CK-NN classifier

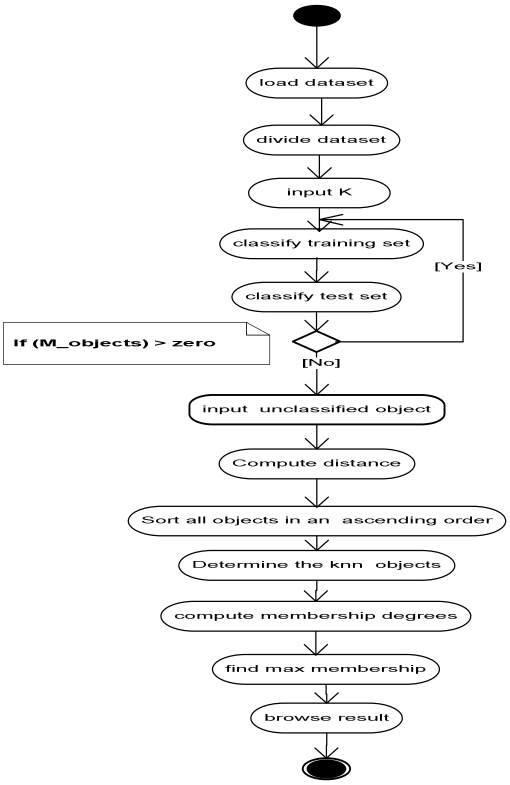

Figure-4: Activity diagram of FK-NN classifier.

The class diagram is used to describe the essential attributes and operation of the system classes and explain the relationship between them. Figure-5 depicts the class diagram of the CFK-NN system.

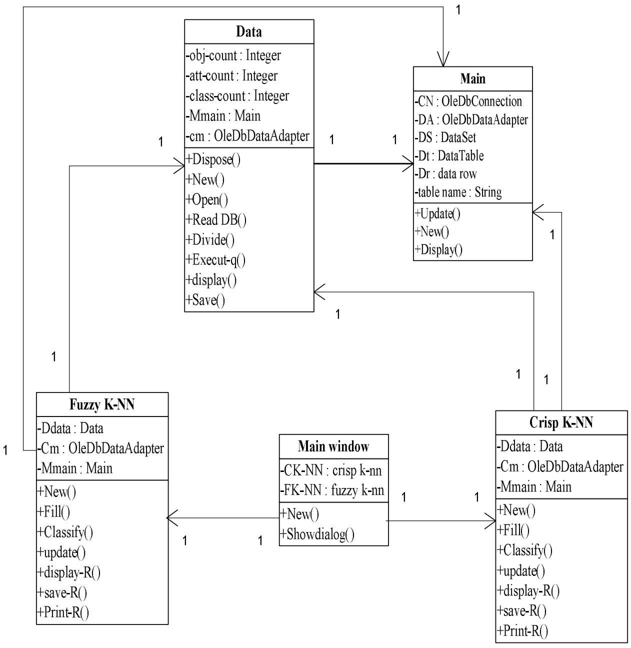

Figure-5: Class diagram for CFK-NN classification system.

The Sequence diagram explains how objects interact with each other in order to accomplish the systems' goals. The Sequence diagram uses three essential elements known as; objects, messages, and object lifeline. It is used to display massages and to trace actions between all objects of the system. Figure-6 depicts the Sequence diagram for CK-NN classifier and figure-7 depicts the Sequence diagram for FK-NN classifier.

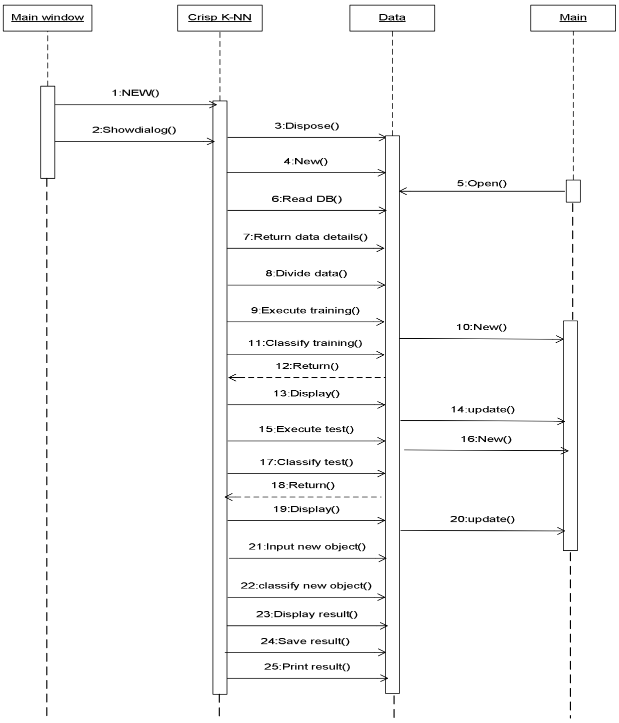

Figure-6: Sequence diagram for CK-NN classifier sub-system.

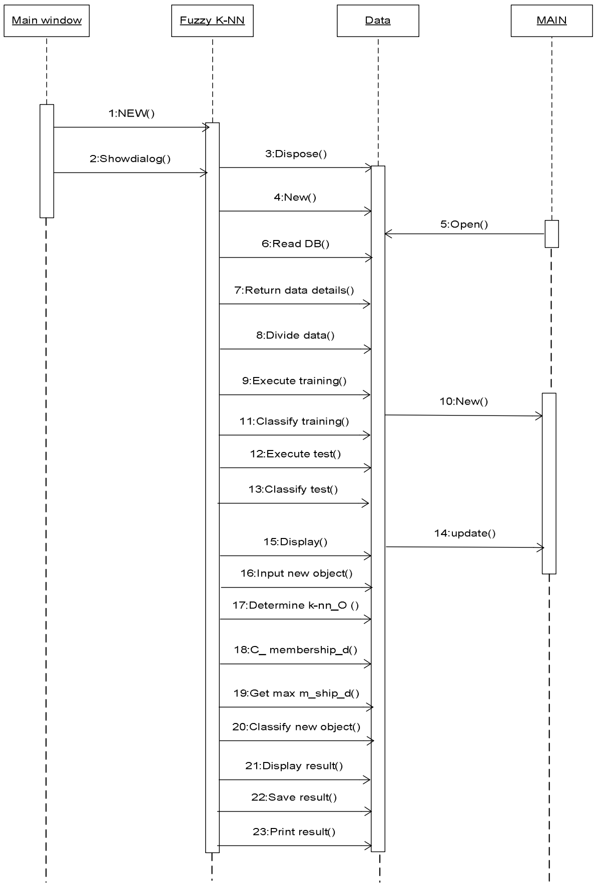

Figure-7: Sequence diagram for FK-NN classifier sub-system.

5.0 EXPERIMENTS AND RESULTS

Here, we demonstrate the obtained results from the experiments conducted on our system. We compare the obtained results with the original result of the used data sets. The system performance is tested with two real data sets. All the experiments are performed on a Laptop computer with 1.73GHz Intel processor, 504 MB of RAM memory, 80 GB Hard disk and Microsoft Windows XP operating system. Our system is tested and evaluated by running two experiments. In testing and evaluating our system, we define the maximum value for K to be the number of classes plus 1 in order to break the ties. The purpose of these experiments is to evaluate the effectiveness of our classifier in the validation phase of the system. Criteria

5.1 FIRST EXPERIMENT

The first experiment is conducted on a data set for fitting contact lenses that consists of 24 records each of which is described by six attributes. We have tested our system with different values of K from 1 to 4 and each time the experiment is run 5 times. Table-1 lists the final obtained results of the 20 runs with different values of K. In each run the M_records (misclassified records) are noted.

Table-2 lists the average accuracy of the classifier of different values of K obtained from 20 runs of the system.

Table-1: The classifier accuracy results of the 20 runs.

Test Run	K Values	M Records	Error Rate	Classification Accuracy
1	1	2	0.222	0.778
2	2	5	0.556	0.444
3	3	2	0.222	0.778
4	4	2	0.222	0.778
5	1	0	0.000	1.000
6	2	5	0.556	0.444
7	3	2	0.222	0.778
8	4	0	0.000	1.000
9	1	4	0.500	0.500
10	2	5	0.625	0.375
11	3	1	0.125	0.875
12	4	1	0.125	0.875
13	1	4	0.500	0.500
14	2	4	0.500	0.500
15	3	1	0.125	0.875
16	4	1	0.125	0.875
17	1	2	0.222	0.778
18	2	4	0.500	0.500
19	3	1	0.125	0.875
20	4	1	0.125	0.875

Table2: The average accuracy of classifier with different values of K.

Number of neighbors (K)	Average accuracy
1	0.706
2	0.453
3	0.831
4	0.881
Over all average accuracy	0.719

Interpretation of the results is that the best accuracy of the classifier is 88.1% at a value of K equal to 4 of this data set. The over all average accuracy of the classifier of the 20 runs is 71.9%

We compared our classifier's results with the original classification results. Hence we found that there are only three cases that are incorrectly classified. So, we can say that our system has given an excellent result for this type of data with accuracy of 87.5% when K=4 in both of the subsystems (CK-NN classifier and FK-NN classifier).

Interpretation of the results is that the best accuracy of the classifier is 88.1% at a value of K equal to 4 of this data set. The over all average accuracy of the classifier of the 20 runs is 71.9%

We compared our classifier's results with the original classification results. Hence we found that there are only three cases that are incorrectly classified. So, we can say that our system has given an excellent result for this type of data with accuracy of 87.5% when K=4 in both of the subsystems (CK-NN classifier and FK-NN classifier).

5.2 SECOND EXPERIMENT

The second experiment is carried out on the Iris dataset that consists of 150 records with 3 classes, each class has 50 records. The 3 classes are: Iris setosa, Iris versicolor and Iris virginica. Each record is described by six attributes. The effectiveness of our classifier is evaluated by comparing our results against the original Iris data results as a validation phase. Table-3 lists the final obtained results when the experiment is run twenty times to get the best results possible with different values of K. Table-4 lists the average accuracy of the classifier of different values of K obtained from 20 runs of the system.

Table-3: The CK-NN classifier results for second experiment of the Iris data set.

Test runs	K Value	M Records	Error Rate	Classifier Accuracy
1	1	3	0.059	0.941
2	2	8	0.157	0.843
3	3	4	0.078	0.922
4	4	4	0.078	0.922
5	1	5	0.100	0.900
6	2	5	0.100	0.900
7	3	4	0.080	0.920
8	4	4	0.080	0.920
9	1	2	0.039	0.961
10	2	2	0.039	0.961
11	3	1	0.020	0.980
12	4	1	0.020	0.980
13	1	2	0.039	0.961
14	2	3	0.059	0.941
15	3	2	0.039	0.961
16	4	2	0.039	0.961
17	1	4	0.080	0.920
18	2	4	0.080	0.920
19	3	3	0.060	0.940
20	4	3	0.060	0.940

Table-4: The results of averages of classifier accuracys.

Number of neighbors (K)	Average accuracy
1	0.937
2	0.913
3	0.945
4	0.945
Over all average accuracy	0.935

Interpretation of the results is that the best accuracy of the classifier is 94.5% at the values of K equal to 3 and 4 of this data. The over all average accuracy of the classifier of the 20 runs is 93.5%. In comparing the results of the two subsystems and the original results of the Iris data set, we have found that there are only three cases that are incorrectly classified. So we can say that our subsystem CK-NN has given an excellent result for this set of data with classification accuracy of 98% when K=4 and K=3 and the FK-NN classifier has given an excellent result for this data set with classification accuracy of 100% when K=4 and K=3.

6.0 CONCLUSION

We have achieved the objectives of this work via the implementation of the Crisp K-Nearest Neighbor algorithm as CK-NN classifier and Fuzzy K-Nearest Neighbor algorithm as FK-NN classifier. We had conducted two experiments each of the consisted of 20 runs. Table-5 summarizes our comparison results of the experiments.

Table-5: Summarization of experiments results.

	First experiment	Second experiment
Data set name	Fitting contact lenses dataset	The iris dataset
Number of records	24 records	150 records
Number of attribute	6 attribute	6 attribute
Class name	Class	Species name
No of run	20 runs	20 runs
Best value of K	K= 4	K=3, K= 4
The average over all accuracy	71.9%	93.5%

We would like to comment on the differences in the results that could be due to:

1. To the method used in splitting the data set into training and testing sets. Where we have spilt the data randomly into two-thirds and one-third as training test and test set respectively.
2. To the different value of K.
3. To the method used to implementations of the algorithms.
4. To the programming languages and operating system used.
5. To the distance function used to compute the distances between objects.

From all of the experiments that we had carried out on our system, we can conclude that there is an advantage of the FK-NN classifier than the CK-NN classifier in the sense that the FK-NN classifier had given the correct classification when the CK-NN classifier had given a tie for some of the cases.

7.0 FURTHER RESEARCH

1. To perform more experiments with our system with different size of databases.
2. To use other re-sampling methods of dividing the available dataset into training set and testing set.
3. To use another distance function to compute the similarity between the objects in the data set.
4. To add some modification to this system to use with other predictive tasks such as estimation or prediction tasks.

REFERENCES

[1] Alpaydin, E., *Introduction to Machine Learning*, MIT Press. United States of America, 2004.

[2] Bechtel, W. and Herschbach, M., Philosophy of the Cognitive Sciences, SUNY Press, Albany, 2008.

[3] Berry, M. and Linoff, G., Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management, Wiley Publishing, Inc., Indiana, USA, 2004.

[4] Bezdek, J., Keller, J., Krisnapuram, R. and Pal, N., Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Springer Science and Business Media, Inc., United States of America, 2005.

[5] Cornell, G. and Morrison, J., Programming VB.NET: A Guide for Experienced Programmers, Springer, United States of America, 2002.

[6] Fayyad, U., Piatetsky, G., and Smyth, P., From Data Mining to Knowledge Discovery in Databases, American association for Artificial Intelligence, United States of America, 1996.

[7] Fowler, M. and Scott, K., UML Distilled A Brief Guide to the Standard Object Modeling Language, Addison Wesley, USA, 2000.

[8] Han, J. and Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann publishers, Canada, 2000.

[9] Hand, D., Mannila, H., and Smyth, P., Principles of Data Mining, Massachusetts Institute of Technology Press, England, 2001.

[10] Huellermeier, E., Fuzzy Methods in Machine Learning and Data Mining: Status and Prospects, Hindawi Publishing Corporation, Los Angeles, 2005.

[11] Joolingen, W., Cognitive tools for discovery learning, "International Journal of Artificial Intelligence in Education", Vol. 10, pp 385 – 397, 1999.

[12] Kantardzic, M., Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley and Sons, Inc, United States of America, 2003.

[13] Larose, D., Discovering knowledge in data: An introduction to Data Mining, John Wiley & Sons, Inc, United States of America, 2005.

[14] Merz, C., and Murphy, P., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 1996.

[15] Mitra, S. and Charya, T., Data Mining Multimedia, Soft Computing, and Bioinformatics, John Wiley & Sons, Inc, United States of America, 2003.

[16] Pender, T., UML Weekend Crash Course, Wiley Publishing Inc, Indiana, USA, 2002.

[17] Ross, T., Fuzzy Logic Engineering Applications, McGraw-Hill, New York, 1995.

[18] Smilkstein, R., The Natural Human Learning Process Guidelines for Curriculum Development and Lesson Plans, Corwin Press, Washington, USA, 2003.

[19] Wang, L. and Xiuju, F., Data Mining with Computational Intelligence, Springer, Germany, 2005.

[20] Witten, I. and Frank, E., Data Mining: Practical Machine Learning Tools and Technique, Elsevier Inc, United States of America, 2005.