A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset

Y. Angeline Christobel, P.Sivaprakasam

***Abstract** - The general problem for data quality is missing data. The real datasets have lot of missing values. Mean method of imputation is the most common method to replace the missing values. In our previous work [23], we address the negative impact of missing value imputation and solution for improvement while evaluating the performance of kNN algorithm for classification of Diabetes data. In this paper, we address a new Class-wise k Nearest Neighbor (CkNN) method for the Classification of Diabetes Dataset. We selected diabetes dataset because it contains lot of missing values and the impact of imputation is very obvious. To measure the performance, we used Accuracy, Sensitivity and Specificity and Error rate as the metrics. The arrived results show the significant improvement measured with respect to the above metrics.*

Key Words: *Data Mining, Classification, kNN, Imputation, Data Normalization and Scaling.*

I. INTRODUCTION

The k-nearest neighbor (kNN) approach has effectively been used in different data analysis applications [5,6] such as information retrieval, database, pattern recognition, data mining and machine learning due to its simplicity, easy-understanding and fairly high accuracy. It has recently been recognized as one of the top 10 algorithms in data mining [7]. In data mining, KNN approach is mainly used in clustering and classification.

In the machine learning, lot of research work has been done to solve the classification problem. Pima Indian Diabetes dataset is very difficult to classify due to its missing values. Lot of research has been done on Pima Indian Diabetes dataset to improve the classification accuracy. In [8], Michie, Spiegelhalter and Taylor used various methods to classify Pima Indians Diabetes dataset. They calculated miss classification accuracies for different machine learning algorithms. It is shown in Table 1.

Table.1 Error rate on Pima Indian Diabetes dataset

S.No	Algorithm	Error rate(%)
1	Discrim	22.50
2	Quadisc	26.20
3.	Logdisc	22.30
4	SMART	23.20
5	ALLOC80	30.10
6	KNN	32.40
7	CASTLE	25.80
8	CART	25.50
9	IndCART	27.10
10	NewID	28.90
11	AC2	27.60
12	Baytree	27.60
13	NaiveBay	26.20
14	CN2	28.90
15	C4.5	27.00
16	Itrule	24.50
17	Cal5	25.00
18	Kohonen	27.30
19	DIPOL92	22.40
20	Backprob	24.80
21	RBF	24.30
22	LVQ	27.20

The error rate of KNN is 32.4%. The average miss classification rate is 26.20%. This is the standard error rate for different machine learning algorithms on Pima diabetes dataset.

Jeatrakul and Wong in [12] have done a comparative study on five different types of neural network architectures such as Back propagation Neural Network (BPNN), Radial basis function Neural Network (RBFNN), General Regression Neural Network (GRNN), Probabilistic Neural Network (PNN) and Complementary Neural Network (CMTNN) for pima Indian diabetes dataset. Table. 2 shows the error rate% for different neural network architecture.

Table.2 Error rate for different neural network architecture

S.No	Neural Network Type	Error rate(%)
BPNN	BPNN	23.83
2	GRNN	24.74
3	RBFNN	23.44
4	PNN	24.74
5.	CMTNN	23.51

Manuscript received on February 13.

Y.Angeline Christobel, Research Scholar, Department of Computer Science, Karpagam University, Coimbatore, India.

Dr. P. Sivaprakasam, Professor, Department of Computer Science, Sri Vasavi College, Erode, India.

The average error rate is 24.05%. In [11], Estebanez, Alter and Valls used genetic programming for classification tasks. They reduced the input dimension from 8 to 3. They applied SVM(Support Vector Machine), Simple logistics and Multilayer Perceptron algorithms on Pima Indian diabetes data for classification. The error rate for SVM is 22%, Simple Logistics is 22.14% and Multilayer perceptron is 23.31%.

Misra and Dehuri in [10] created a Functional link Artificial Neural network and compared its classification performance with other machine learning algorithms. The error rate of their FLANN is 21.87% , MLP - 24.8%, KNN – 30.3, CART- 25.5, C4.5 – 25.3, FSS – 26.4, BSS – 32.3, MFS1 – 31.5, MFS2 – 27.5.

Twala, Jones, and Hand [13] proposed a method for creating a separate class for missing values, and found that its performance was competitive with that of likelihood-based multiple imputation.

In [14], D. Vijayalaksmi and K. Thilagavathi proposed a clustering algorithm based on a graph b-colouring technique for Pima Indian Diabetic dataset. They compared their algorithm with KNN and K-means clustering. In terms of accuracy and purity, the clustering based on graph colouring outperforms the KNN and K-means method.

A Satheesh et al [15] designed a dynamic nearest neighbor classifier for data integrated via object oriented concept and generalization.They integrated the data collected from multiple service providers into a single consolidated unit(training instance). They enhanced the traditional KNN using normalization and majority voting. The results show that the dynamic nearest neighbor is more efficient when compared to the traditional KNN.

The objective of this paper is to address the ways to improve the classification of the kNN algorithm for classification of Diabetes data with imputed missing values. We propose a new class wise nearest neighbor algorithm (CkNN) for improving the performance of the standard kNN.

II. DATA PREPROCESSING

The data preprocessing techniques have a significant impact on the performance of machine learning algorithms. To produce quality mining results, data preprocessing is very important. The challenging issue in machine learning and data mining is missing values imputation [20]. Good database design and analysis can reduce the missing data problems. The right technique should be selected to handle missing values depending on problem domain and the goal.

In this paper, we used mean substitution to impute missing values and data scaling algorithm to increase the accuracy of KNN and also propose a class wise nearest neighbor algorithm(CKNN) to improve the performance of KNN. These algorithms are discussed below:

A. Mean substitution

The popular imputation method to fill in the missing data values is to use a variable's mean or median.

The following algorithm explains the very commonly used form of mean substitution method [18]

Let
 D = { A1, A2, A3, ….. An }
Where
D is the set of data with missing values
Ai – is the ith attribute column of values of D with missing values in some or all columns
 n - is the number of attributes.

```
Function MeanSubstitution(D)
Begin
  For i=1 to n {
    ai ← Ai ∩ mi
where
ai is the column of attributes without missing values
 mi is the set of missing values in Ai (missing values denoted by a symbol)
    Let μi be the mean of ai
Replace all the missing elements of Ai with μi
  }
  Finally we will have the imputed data set.
End
```

B. Data Normalization

Normalization is a scaling down transformation of the instances. Within an instance there is often a large difference between the maximum and minimum values. When normalization is performed the value magnitudes and scaled to appreciably low values. This is important for many neural network and KNN algorithms.

The Simple Data Scaling Algorithm

The following algorithm explains the data scaling method[19].

Let
 D = { A1, A2, A3, ….. An }
Where
 D is the set of unnormalized data
Ai – is the ith attribute column of values of
m- is the member of rows (records)
 n - is the number of attributes.

```
Function Normalize(D)
Begin
  For i=1 to n {
    Maxi ←max(Ai)
    Mini←min(Ai)
    For r =1 to m {
      Air ← Air-Mini
      Air ← Air/ Maxi
      Where
Air is the element of Ai at row r
}
  }
  Finally we will have the scaled data set.
End
```

C. k-Nearest Neighbor(knn) Classification Algorithm

KNN classification classifies instances based on their similarity. It is one of the most popular algorithms for pattern recognition. It is a type of Lazy learning where the function is only approximated locally and all computation is deferred until classification.

An object is classified by a majority of its neighbors. K is always a positive integer. The neighbors are selected from a set of objects for which the correct classification is known.

The Pseudo Code of kNN Algorithm

The following shows the pseudo code of KNN algorithm

```
Function kNN(train_patterns , train_targets, test_patterns )
Begin
```

```
Uc   - a set of unique labels of train_targets;
N   -  size of test_patterns
for i = 1:N{
dist:=EuclideanDistance(train_patterns,  test_patterns(i))
idxs := sort(dist)
topkClasses := train_targets(idxs(1:Knn))
c := DominatingClass (topkClasses)
test_targets(i)  :=  c
}
End
```

D. The proposed CKNN Algorithm

The following shows the pseudo code of CKNN algorithm

The Pseudo Code of CkNN Algorithm

```
Function CkNN(train_patterns , train_targets, test_patterns )
Begin
Uc   - a set of unique labels of train_targets;
N   -  size of test_patterns
C – size of Uc (number of classes)
for i = 1:N{
      for j = 1 to C {
    idx= find(train_targets == Uc(j));
dist:=EuclideanDistance(train_patterns(*,idx),
test_patterns(*,i))
    [d, idxs] := sort(dist)
     md(c)= sum(d (1 to Knn) );
                }
      best = min(md);
                    test_targets(i)  :=  Uc(best)
         }
end
```

In the above algorithm, the first for loop is repeated for N number of records in the testing set and the inner for loop is repeated N times for each class in the training data.

For each class in the training set, the class-wise distance with a testing data record is calculated. That particular testing data is classified to a class label corresponding to the lowest distance. This is repeated for all the records in the testing data set.

E. Validating the Performance of The Classification Algorithm

Classifier performance depends on the characteristics of the data to be classified. Performance of the selected algorithms is measured for Accuracy and Error rate. In this study, we have selected k-fold cross validation for evaluating the classifiers. In k-fold cross validation, the initial data are randomly partitioned into k mutually exclusive subset or folds d1,d2,…,dk, each approximately equal in size. The training and testing is performed k times. In the first iteration, subsets d2, …, dk collectively serve as the training set in order to obtain a first model, which is tested on d1; the second iteration is trained in subsets d1, d3,…, dk and tested on d2; and so on[21]. The accuracy of the classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data [21].

The Accuracy, Error rate, Sensitivity and Specificity can be defined as follows:

Accuracy = (TP+TN) / (TP + FP + TN + FN)
Error rate = (FP+FN) / (TP + FP + TN + FN)
Sensitivity = TP/(TP + FN)
Specificity = TN /(TN + FP)
Where
TP is the number of True Positives
TN is the number of True Negatives
FP is the number of False Positives
FN is the number of False Negatives

III. THE IMPLEMENTATION AND RESULTS

We used Matlab 6.5 on a Core i7 laptop with 4GB of RAM for implementing the algorithms. We used Pima Indians Diabetes data set with lot of missing values.

Pima Indian Diabetes Data set

The Pima Indian diabetes data set is taken from the UCI machine learning repository [22].The data set has 768 instances with two class problems to test whether the patient is positive or negative for diabetes. The patients in this dataset are Pima Indian Women who lives near Phoenix Arizona, USA. This data set consists of 9 attributes as shown in Table1 [16] [17].

Table 1. Diabetes data set

No.	Attribute	Description	Missing Values
1	pregnant	Number of times pregnant	110
2	glucose	Plasma glucose concentration (glucose tolerance test)	5
3	pressure	Diastolic blood pressure (mm Hg)	35
4	triceps	Triceps skin fold thickness (mm)	227
5	insulin	2-Hour serum insulin (mu U/ml)	374
6	mass	Body mass index (weight in kg/(height in m)^2)	11
7	pedigree	Diabetes pedigree function	0
8	age	Age (years)	0
9	diabetes	Class variable (test for diabetes)	0

Class Distribution: Class value 1 is interpreted as "tested positive for diabetes"
Class Value : 0 - Number of instances - 500
Class Value : 1 - Number of instances – 268

The following 3D plot clearly shows the complex distribution of the positive and negative records in the original Pima Indians Diabetes Database which will complicate the classification task.

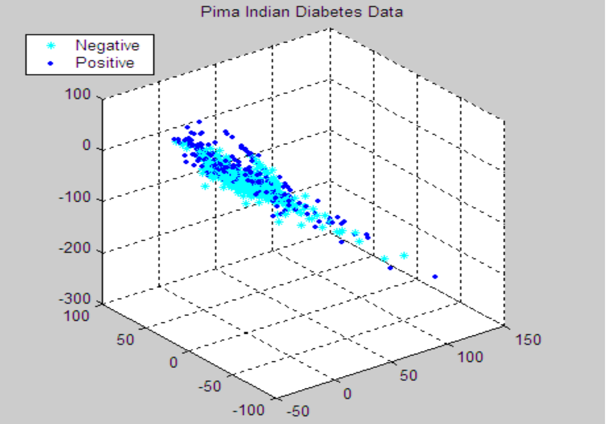

Fig 1. The 3D plot of original Pima Indians Diabetes Database

Results of 10 Fold Validations

We chose to use the results of the 10 fold validation in this evaluation. The following table shows the results of 10 fold validations in terms of sensitivity, specificity, accuracy, error and run time.

Table 1. The Results 10 Fold Validation

Algorithm	Sensitivity	Specificity	Accuracy	Error	Time
kNN	57.34	79.88	71.84	28.16	0.17
CkNN	61.84	87.38	78.16	21.84	0.19

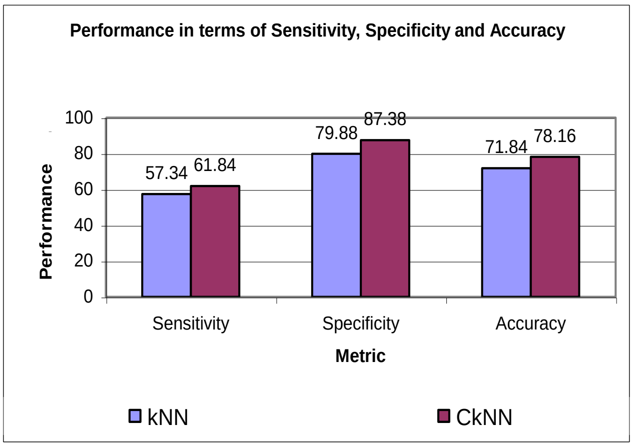

Fig 2. Comparison of Performance –10 fold Validation

The Performance of classification measured with respect to sensitivity, specificity and accuracy has been increased significantly in the case of proposed CkNN algorithm. The above graph clearly shows the increase in performance. Normally, during classification, if sensitivity is increased then it may reduce specificity; but with the proposed CkNN algorithm, the performance is increased with respect to sensitivity, specificity and accuracy.

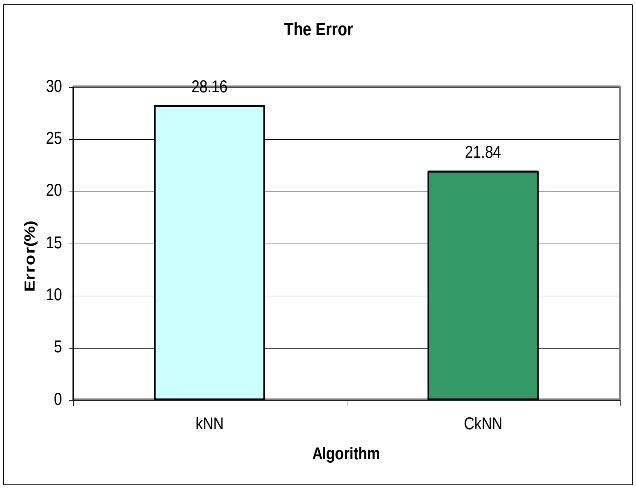

Fig 3. Comparison of Error –10 fold Validation

The classification error has been significantly lower in the case of proposed CkNN algorithm. The above graph Fig 3 clearly shows the decrease in error.

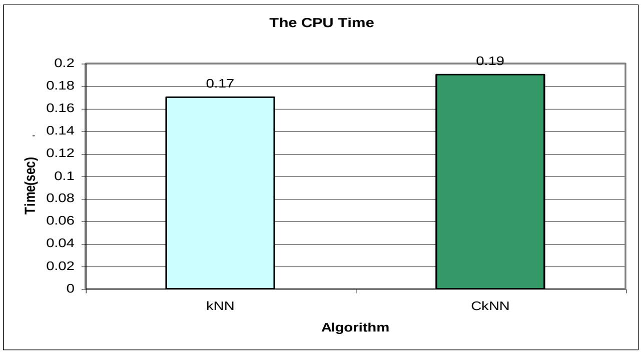

Fig 4. Comparison of CPU time – 10 fold Validation

There is not much increase in CPU time in the case of proposed CkNN algorithm. The difference in CPU time is negligible and will not have much impact on practical applications.

IV. THE CONCLUSION AND FUTURE WORK

We have successfully implemented and evaluated the proposed CkNN classification algorithm and measured the average performance of the algorithm by considering the k fold cross validation with k=10. We compared the performance of the CkNN algorithm with standard kNN.

The Performance of classification measured with respect to sensitivity, specificity and accuracy has been increased significantly in the case of proposed CkNN algorithm. The classification error has been significantly reduced in the case of proposed CkNN algorithm. The CkNN algorithm consumed almost equal CPU time like kNN. The small difference in CPU time is negligible and will not have much impact on practical applications.

Future works may address hybrid classification model using kNN with other techniques.

V. REFERENCES

1. Roshawnna Scales, Mark Embrechts, "Computational intelligence techniques for medical diagnostics".
2. World Health Organization. Available: http://www.who.int
3. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
4. Denoeux, T. 1995: A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics:, 25, 804–813.
5. E. Blanzieri and F. Melgan (2008). Nearest Neighbor Classification of Remote Sensing Images With the Maximal Margin Principle. IEEE Trans.Geoscience and Remote Sensing, 46(6): 1804-1811.
6. Chen, J., and Shao, J., (2001). Jackknife variance estimation for nearest-neighbor imputation. J. Amer. Statist. Assoc. 2001, Vol. 96: 260-269.
7. Wu, XD., et al. (2008). Top 10 Algorithms in Data Mining, Knowledge and Information Systems, 14(1): 1-37.
8. Michie.D, Spiegelhalter. D.J and Taylor .C.C. Machine Learning, Neural and Statistical Classification, Chapter 9 page No 157,158.
9. Lena Kallin Westin, Missing data and the preprocessing perceptron, page 3,ISSN-0348-0542.
10. Misra .B.B, Dehuri. S, 2007, Functional Link Artificial Neural Network for Classification Task in Data Mining, Journal of Computer Science 3 (12):948-955, ISSN 1549-3636 Science Publications.
11. Estebanez .C, Aler .R and Valls. M, Genetic Programming Base Data Projections for Classification Tasks, World Academy of Science, Engineering and technology 2005, Pages 56-61.
12. Jeatrakul .P and Wong .W.K, Comparing the Performance of Different Neural Networks for Binary Classification Problems, 2009 Eighth International Symbosium on Natural Language Processing, Page 111-115.
13. Twala, M.C. Jones, and D.J. Hand." Good methods for coping with

missing data in decision trees”,Pattern Recognition Letters, 29:950–956, 2008.

14. D. Vijayalaksmi, K, Thilagavathi, “ An approach for prediction of Diabetic Siseases by Using b-colouring Technique in Clustering Analysis” International Journal of Applied Mathematical Research, 1 (4) (2012) 520-530
15. Ajita Satheesh, Ravindra Patel, “Dynamic Nearest Neighbours Classifier For Integrated Data Using Object Oriented Concept Generalization” IJSSST, Vol.11, No. 1
16. H. Temurtas, N. Yumusak, and F. Temurtas, “A comparative study on diabetes disease diagnosis using neural network: Expert System with Applications”, 36, 2009, 8610-8615.
17. A. J. Seibel, “Diabetes Guide WebMD”, http://diabetese.webmd.com/guide/oral-glucose-tolerance-test, 2007
18. R.S. Somasundaram and R. Nedunchezhian, "Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", International Journal of Computer Applications Issn-09758887, 2011
19. Angeline Christobel. Y, P. Sivaprakasam “Improving the Performance of K-Nearest Neighbor Algorithm for the Classification of Diabetes Dataset with missing values”, International Journal of Computer Engineering and Technology” (IJCET), Volume 3, Issue 3, October - December (2012), pp. 155-167
20. Brian D. Ripley (1996), Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge.
21. J. Han and M. Kamber,”Data Mining Concepts and Techniques”, Morgan Kauffman Publishers, USA,
2006.
22. UCI Machine Learning Repository
[http://archive.ics.uci.edu/ml/datasets]
23. Angeline Christobel.Y, P.Sivaprakasam,” The Negative Impact of Missing Value Imputation in Classification of Diabetes Dataset and Solution for Improvement”, IOSRJCE, Volume 7, Issue 4(Nov-Dec).pp 16-23
24. R. J. Little and D. B. Rubin. “Statistical Analysis With Missing Data”, Hoboken, NJ: Wiley,(1987).
25. Gustavo E. A. P. A. Batista and Maria Carolina Monard, An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Applied Artificial Intelligence 17(5-6): 519-533 , 2003
26. Alireza Farhangfar, Lukasz Kurgan, Jennifer Dy,”Impact of imputation of missing values on classification error for discrete data”, Pattern Recognition, Volume 41, Issue 12, December 2008, Pages 3692–3705
27. Qinbo Song, Martin Shepperd.”A new imputation method for small software project data sets”, Journal of Systems and Software, Volume 80, Issue 1, January 2007, Pages 51–62
28. José M. Jerez Ignacio Molina ,, Pedro J. García-Laencina ,Emilio Alba, Nuria Ribelles, Miguel Martín, Leonardo Franco,” Missing data imputation using statistical and machine learning methods in a real breast cancer problem”, Artificial Intelligence in Medicine, Volume 50, Issue 2, October 2010, Pages 105– 115
29. Vidan Fathi Ghoneim, Nahed H. Solouma, Yasser M. Kadah,” Evaluation of Missing Values Imputation Methods in cDNA Microarrays Based on Classification Accuracy”, 978-1-4244-7000-6/11 IEEE.
30. K.T. Chuang, K. P. Lin, and M. S. Chen. “Quality-Aware Sampling and Its Applications in Incremental Data Mining”, IEEE Transactions on knowledge and data engineering,vol.19,no. 4,2007.
31. Li.Liu, Y. Tu, Y. Li, and G. Zou. “Imputation for missing data and variance estimation whenauxiliary information is incomplete”, Model Assisted Statistics and Applications, 83-94,2005
32. Y Shi, Z Cai, G Lin, Classification accuracy based microarray missing value imputation. in Bioinformatics Algorithms: Techniques and Applications, ed. by Mandoiu I, Zelikovsky A (Wiley-Interscience, Hoboken, NJ, USA, 2007), pp. 303–328
33. J Hua, T Waibhav, ER Dougherty, Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition 42(3), 409–424 (2009).
34. Brian A. Cattle1, Paul D. Baxter, Darren C. Greenwood, Christopher P. Gale1, Robert M. West, “Multiple imputation for completion of a national clinical audit dataset”, Statistics in Medicine Volume 30, Issue 22, pages 2736–2753, 30 September 2011
35. Awawtam. "Pima Stories of the Beginning of the World." The Norton Anthology of American Literature. 7th ed. Vol. A. New York: W. W. Norton &, 2007. 22-31
36. Angeline Christobel, SivaPrakasam,”An Empirical Comparison of Data mining Classification Methods”, International Journal of Computer Information Systems, Vol. 3, No. 2, 2011