International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 11-15 www.ijaem.net ISSN: 2395-5252

An Efficient K Nearest Neighbor Imputation Method For Missing Values

[1,]M. Gobi, [2,]G. Buvaanyaa
[1]*Assistant Professor, Department of Computer Science, Chikkanna Government Arts College, Tirupur, Tamilnadu, India*
[2]*M.Phil., Research Scholar, Department of Computer Science, Chikkanna Government Arts College, Tirupur, Tamilnadu, India*
Corresponding Author: G. Buvaanyaa

Date of Submission: 07-07-2020 Date of Acceptance: 21-07-2020

ABSTRACT: Classification is a data mining technique used to predict group membership for data instances within a given dataset. It is used for classifying data into different classes by considering some constrains. Classification is considered as an example of supervised learning as training data associated with class labels is given as input. Predictive classification as a wide range of application in data mining. Most real data set have missing values which affects the accuracy of classifiers. This paper will investigate predictive performance of missing data using KNN classifier.

KEYWORDS: Classification, Data Mining, Classification Technique, K- NN classifier, Predictive.

I. INTRODUCTION

Data mining is most important analysis step of knowledge discovery in database (KDD) process. The main goal of data mining is to extract the useful information from huge raw data and converting it to an understandable form for its effective and efficient use. In common, data mining tasks can be divided into two categories: descriptive and predictive classification techniques. Classification algorithm assigns each instance to a particular class such that classification error will be least. It is used to extract models that accurately define important data classes within the given dataset.

Data Mining has a lot of functions, such as description, association analysis, classification and prediction, clustering analysis etc. Among all, classification and prediction are widely used in many fields. However, in real-world datasets, there are many problems in data quality such as incompleteness, redundancy, inconsistency, noise data etc.

Classification process is divided into two main steps. The first is the training step where the classification model is built. The second is the classification itself, in which the trained model is applied to assign unknown data object to one out of a given set of class label.

This paper focuses on a classification technique that are most commonly used in data mining. This would provide the guideline for interesting research issues which in turn help other researchers in developing innovative algorithms for applications or requirements which are not available.

II. CLASSIFICATION AND PREDICTION

A. CLASSIFIER

Classification means constructing a classifying function or model from the known data. Such function or model can also be called "classifier", which can classify the records in database into given classes, thus can predict the unknown variables under some given conditions. Classifiers differ greatly in prediction accuracy, training time and number of leaves (Decision Trees). There is not a classifier which performs best in all aspects. Prediction accuracy of classifiers can be affected by the factors as follows.

B. NUMBER OF RECORDS IN TRAINING SUBSET

Classifier needs to learn from training set. Therefore, larger training set makes the classifier more reliable. But the training time is also become longer.

C. DATA QUALITY

Problems such as noise data, missing data, data inconsistency etc. bring a lot of wrong information which will lead to wrong classify. It is impossible to build a convictive classifier with incompleteness or wrong data.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 11-15 www.ijaem.net ISSN: 2395-5252

D. ATTRIBUTE QUALITY

Attributes provide information for classifying. The prediction accuracy can be improved by including more attributes. However, more attributes means calculating more attribute combinations and more training time. It is essential to choose attributes which are valuable for classification.

E. CHARACTERISTICS OF THE RECORDS TO BE PREDICTED

If Characteristics of the records to be predicted are different from records in training set, it may lead to high incorrect rate.

III. DATASET

The data set as mentioned below is taken from a Kaggle Repository: Rheumatoid Arthritis dataset. It has total 996 instances and 7 attributes likeI'd, Year, Gender, Age, Treatment, Baseline, Time.The attributes are real and of multivariate characteristics. This data was first converted into CSV file using JSON file from the website using Python. From the 996 instances we have here discussed about 30 instances. In that we are knowing about, the knowedge of how the K Nearest Neighbor algorithm works.

TABLE 1: RHEUMATOID ARTHRITIS DATASET

S.NO	I'D	YEAR	SEX	AGE	TREAT MENT	BASE LINE	TIME
1	1	4	2	54	2	2	1
2	1	5	2	54	2	2	3
3	1	5	2	54	2	2	5
4	2	4	1	41	1	3	1
5	2	4	1	41	1	3	3
6	2	4	1	41	1	3	5
7	3	3	2	48	2	3	1
8	3	4	2	48	2	3	3
9	3	4	2	48	2	3	5
10	4	4	2	40	1	3	1
11	4	3	2	40	1	3	3
12	4	4	2	40	1	3	5
13	5	4	2	29	2	3	1
14	5	3	2	29	2	3	3
15	5	4	2	29	2	3	5
16	6	3	2	43	2	2	1
17	6	2	2	43	2	2	3
18	6	3	2	43	2	2	5
19	7	3	2	55	2	4	1
20	7	2	2	55	2	4	3
21	7	3	2	55	2	4	5
22	8	3	1	49	1	3	1
23	8	2	1	49	1	3	3
24	8	2	1	49	1	3	5
25	9	2	2	64	1	2	1
26	9	2	2	64	1	2	3
27	9	1	2	64	1	2	5
28	10	3	2	51	2	3	1
29	10	3	2	51	2	3	3
30	10	4	2	51	2	3	5
31	11	4	2	46	2	4	?

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 11-15 www.ijaem.net ISSN: 2395-5252

The Dataset which belongs to Rheumatoid Arthritis category, having a question whether I'D number 11 belongs to which Time period. So, that we have to predict the result using KNN Classifier. For that, We have to find the distance between all of these to find the result.

The Distance equation is normally,

$$\sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

i.e., Euclidean Distance method

Other distances are, Manhattan Distance and Minkowski Distance etc.,

Using this Euclidean Distance method we are having an Distance Table of every patient I'D, has shown below,

TABLE 2: RHEUMATOID ARTHRITIS DATASET DISTANCE EVALUATED

S.NO	I'D	AGE	SEX	DISTANCE	TIME
1	1	54	2	8	1
2	2	41	1	5.09	3
3	3	48	2	2	5
4	4	40	2	6	3
5	5	29	2	17	5
6	6	43	2	3	1
7	7	55	2	9	3
8	8	49	1	3.16	5
9	9	64	2	2	1
10	10	51	2	5	5

IV. K- NEAREST NEIGHBOR CLASSIFICATION

The K-Nearest Neighbor Algorithm is the simplest of all machine learning algorithms. K-Nearest Neighbor is instance based learning method. Lazy-learning algorithms require less computation time during the training phase than eager-learning algorithms but more computation time during the classification process.

Nearest-neighbor classifiers are based on learning by resemblance, i.e. by comparing a given test sample with the available training samples which are similar to it. For a data sample X to be classified, its K-nearest neighbors are searched and then X is assigned to class label to which majority of its neighbors belongs to Dataset. The choice of k also affects the performance of k-nearest neighbor algorithm. If the value of k is too small, then K-NN classifier may be vulnerable to over fitting because of noise present in the training dataset. On the other hand, if k is too large, the nearest-neighbor classifier may misclassify the test sample because its list of nearest neighbors may contain some data points that are located far away from its neighborhood.

K-NN fundamentally works on the belief that the data is connected in a feature space. Hence, all the points are considered in order, to find out the distance among the data points. Euclidian distance or Hamming distance is used according to the data type of data classes used. In this a single value of K is given which is used to find the total number of nearest neighbors that determine the class label for unknown sample. If the value of K=1, then it is called as nearest neighbor classification.

The K-NN classifier works as follows:

1. Initialize value of K.
2. Calculate distance between input sampleand training samples.
3. Sort the distances.
4. Take top K- nearest neighbors.
5. Apply simple majority.
6. Predict class label with more neighbors for input sample

Following example shows that there are three classes X, Y and Z as shown in figure 1. Now, it is required to find out the class label for data sample P. Here, value of K=3 and the Euclidean distance is calculated for each sample pair and it is found that three nearest neighbor samples are falling in the class label X, while single tuple belongs to class label Z. So, the sample P is assigned to class X as it is the principal class for that sample.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 11-15 www.ijaem.net ISSN: 2395-5252

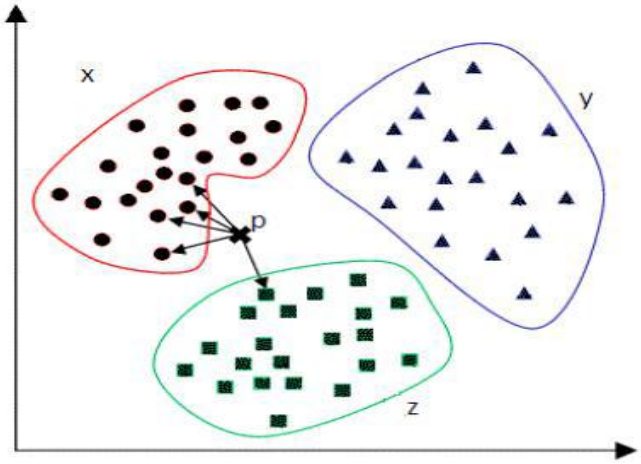

FIGURE 1: AN EXAMPLE OF K-NN CLASSIFIER

DESIGN OF EXPERIMENTS

This paper has selected a classifier to study the influence of missing data to classify, that is, K-Nearest Neighbours classifier (KNN) datasets were collected from Kaggle repository and, are used in the experiments, as shown in Table 1. Three indexes are used to evaluate the missing influence on classifier: Prediction accuracy (Pa), Prediction profit (Pp) and Prediction losing (Pl), which are defined as follows.

$Pa=numberofcorrectlypredictedrecord/numberofallrecords\times 100\%$

$Pp=Pa-MD/MD\times 100\%$

$Pi=AC-Paundercertainmissingrate/AC\times 100\%$

AC is the prediction accuracy without missing data.

MD is the proportion of the class in the dataset which having the greatest number of records.

Without any prediction model, if all the records are classified into that class, MD will be the prediction accuracy of the dataset. Prediction profit is proposed by Peng Liu, Elia El-Darzi et al. (2004) to evaluate the performance of the classifier.

Then, a given percentage, 10%, 20%, … , 80% of missing data is artificially inserted into the training subsets at completely random. Finally, the classification algorithm mentioned above are applied into the training subset to build up classifier, and, these classifier are used to classify instances in testing subset to investigate the classification accuracy.

V. RESULTS AND ANALYSIS

From the above Table 1 and Table 2, having a Dataset and its Distance of Rheumatoid Arthritis category. In that we have get a result of Time period as 1. Because the Priority of the Distance has predicted as 1 so the result is predicted as time Period 1.

In general, when the proportion of missing data in the dataset is less than 10%, they have little adverse impact on the classifiers. If the missing rate is between 10% and 20%, the impact should not be neglected. The average Prediction losing rises to 4.63%. However, the adverse impact can be reduced significantly by some simple methods, such as replacing missing data by an approximation. If the missing rate exceeds 20%, there is an obvious decrease in the prediction accuracy and the missing data should be handled with high cautiousness.

Appropriate methods should be chosen to eliminate the adverse impact of the missing data and optimize the performance of classifiers. In the real world, there are great quantities of missing data in databases and, usually, the proportion of missing data exceeds 20%. However, if the proportion of missing data exceeds 50%, the average prediction losing rises to more than 10%.

Obviously, the loss of prediction accuracy caused by missing data is quite huge. with the increase in the missing rate, the losing of prediction is rising with accelerated paces. That is to say, with the increase of quantity of the missing data, the little raising of the missing rate will result in a larger and larger decrease in the prediction accuracy.

Further more, the pace of KNN is very fast. When missing rate exceeds 10%, there is an

obvious and sharp increase in prediction losing. The K-Nearest Neighbor has the sharpest trend.

VI. CONCLUSION

Missing data may reduce the accuracy of prediction models. This paper mainly studies the impact of missing data to classification algorithms. The sensitivity of classifiers to missing data is analysed. If the proportion of missing data exceeds 20%, there is an obvious decrease in the accuracy of prediction. Methods for missing data treatment should be chosen cautiously to eliminate the negative impact on the classification accuracy and optimize the performance of classifiers. Among the classifier, the K-Nearest Neighbour is the most sensitive.

REFERENCES

[1]. S.Archana and Dr.K.Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Vol. 2 Issue. 2, February 2014.
[2]. BhaveshPatankar and Dr. Vijay Chavda, "A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 12, December 2014.
[3]. Sagar S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms", Oriental Journal of Computer Science & Technology, Vol. 8, April 2015.
[4]. Meenakshi and Geetika, "Survey on Classification Methods using WEKA", International Journal of Computer Applications, Vol. 86, No.18, January 2014.
[5]. H. Bhavsar and A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, Issue. 4, September 2012
[6]. T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification", IEEE Transactions on Information Theory, vol. 13, No. 1, pp. 21-27, 1967.
[7]. J. Han and M. Kamber, "Data Mining Concepts and Techniques", Elevier, 2011.
[8]. K. P. Soman, "Insight into Data Mining Theory and Practice", New Delhi: PHI, 2006.
[9]. S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica, vol. 31, pp. 249-268, 2007.
[10]. M. Soundarya and R. Balakrishnan, "Survey on Classification Techniques in Data mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2014.
[11]. R. Duda, and P. Hart, "Pattern Classification and Scene Analysis", John Wiley and Sons, New York, 1973.
[12]. N. Friedman, D. Geiger, and Goldazmidt, "Bayesian Network Classifiers", Machine Learning, vol. 29, pp. 131-163, 1997.
[13]. "Decision tree learning" pdf.
[14]. [14] Matthew N. Anyanwu and Sajjan G. Shiva, "Comparative Analysis of Serial Decision Tree ClassificationAlgorithms", Researchgate, January 2009.
[15]. Brijain R. Patel and KushikK.Rana, "A Survey on Decision Tree Algorithm for Classification", International Journal of Engineering Development and Research, 2014.

International Journal of Advances in Engineering and Management
ISSN: 2395-5252

IJAEM

Volume: 02 Issue: 01 DOI: 10.35629/5252

www.ijaem.net
Email id: ijaem.paper@gmail.com