Relative study of Prediction KNN Algorithm Using Normalization Techniques

G.Amani[1], K.Venkata Ramana[2]
[1]*Student, Dept of MCA KMM Insitute of Post Graduate Studies, Tirupati*
[2]*K.Venkataramana, Dept of MCA KMM Insitute of Post Graduate Studies, Tirupati*

***Abstract*- The task of classifying a set of documents into different categories from group of sets. Here K- Nearest Neighbors algorithm is used. In this algorithm is is mainly used for data mining and pattern recognition and machine learning because its is very easy to understand and this perfoermance is good. It is non-parametric technique for regression and catagorization. KNN (K- Nearest Neighbors algorithm) is popular method to categorize the dataset. This paper is concerned with the comparative study or analysis of K-Nearest neighbor algorithm under different normalization techniques and different values of K. For the comparative analysis, we have used "IRIS" Dataset. To measure accuracy, Here we are used two normalization techniques that are Z-Score Normalization and Min-Max Normalization. Using these techniques accuracy and performance will be increases compared other techniques using data sets. Also, we have computed the average prediction efficiency of K-nearest neighbor algorithm using the two normalization techniques and concluded the one technique with the highest efficiency.**

***Index Terms*- Data mining, Prediction KNN (K- Nearest Neighbors algorithm), Z-Score and Min-Max Normalization techniques.**

I. INTRODUCTION

In this system, we have different techniques, categorizations, machine learning. In pattern recognition, the k-NN algorithm is a non-parametric method used for regression and classification. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, a useful technique can be to assign weight to the contributions of the neighbors. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data. The algorithm is not to be confused with k-means, another popular machine learning technique.

The technique which provides linear transformation on original range of data is called Min-Mix Normalization. The technique which keeps relationship among original data is called Min-Mix Normalization. Min-Max normalization is a simple technique where the technique can specifically fit the data in a pre-defined boundary with a pre-defined boundary. The technique which gives the normalized values or range of data from the original unstructured data using the concepts like mean and standard deviation then the Parameter is called as Z-score Normalization.

II . RELATED WORK

Knn algorithm is mainly used for data maining and pattern recognition and machine learning because it's very easy to understand and this performance is good. Finance:-KNN as a process of scraping out useful patterns and correlations has its own domain in financial modeling. Stock market forecasting including planning investment strategies, uncovering market trends, what stocks to purchase etc. Are some of the crucial financial tasks of KNN. These further include-Credit rating, Money laundering analysis etc. [1] Hui-Ling Chen, Bo Yang, Gang Wang, Jie Liu, Xin Xu, Su-Jing Wang and Da-You Liu "bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method"

[2] Tam, Kar Yan, and Melody Y. Kiang. "Managerial applications of neural networks: the case of bank failure predictions.

Agriculture:- In practice, KNN is employed less than other data mining techniques in fields related to agriculture. Some applications include simulation of precipitation and other weather parameters [3]Radnaabazar Chinchuluun, Won Suk Lee, Jevin Bhorania and Panos M. Pardalos "Clustering and Classification Algorithms in Food and Agricultural Applications: A Survey"

[4] Reese, Heather, et al. "Applications using estimates of forest parameters derived from satellite and forest inventory data."

III. PREDICTION KNN ALGORITHM

In the k-Nearest Neighbor prediction method, the Training Set is used to predict the value of a variable of interest for each member of a target data set. The structure of the data generally consists of a variable of interest (i.e., amount purchased), and a number of additional predictor variables (age, income, location).

1. For each row (case) in the target data set (the set to be predicted), locate the k closest members (the k nearest neighbors) of the Training Set. A Euclidean Distance measure is used to calculate how close each member of the Training Set is to the target row that is being examined.
2. Find the weighted sum of the variable of interest for the k-nearest neighbors (the weights are the inverse of the distances).
3. 3.XLMiner allows the to selection of a maximum value for k, builds models in parallel on all values of k (up to the maximum specified value), and performs scoring on the best of these models.

Computing time increases as k increases, but the advantage is that higher values of k provide smoothing that reduces vulnerability to noise in the Training Set. Typically, k is in units of tens of units, rather than in hundreds or thousands.

Normalize input data:When this option is selected, the input data is normalized, which means that all data is expressed in terms of standard deviations. This option is available to ensure that the distance measure is not dominated by variables with a large scale. This option is not selected by default.

A)Number of nearest neighbors (k): This is the parameter k in the k-nearest neighbor algorithm. If the number of observations (rows) is less than 50 then the value of k should be between 1 and the total number of observations (rows). If the number of rows is greater than 50, then the value of k should be between 1 and 50. The default value is 1.

When Score on best k between 1 and specified value is selected, XLMiner displays the output for that value. If Score on specified value of k as above is selected, the output is displayed for the specified value of k. The default setting is Score on specified value of k as above.

B)Score Training Data: Select these options to show an assessment of the performance of the tree in classifying the Training Set. The report is displayed according to the specifications are Detailed, Summary, and Lift Chart.

C)Score Validation Data: These options are enabled when a Validation Set exists. Select these options to show an assessment of the performance of the algorithm in classifying the Validation Set. The report is displayed according to the specifications: Detailed, Summary, and Lift Charts.

Score Test Data: These options are enabled when a test set is present. Select these options to show an assessment of the performance of the tree in classifying the test data. The report is displayed according to your specifications: Detailed, Summary, and Lift Charts.

Score New Data: See the Scoring New Data section for information the KNN Stored worksheet.

Partitioning Options: XLMiner V2015 provides the ability to partition a data set from with in a classification or prediction method by selecting Partitioning Options on the *Step 2 of 3* dialog. If this option is selected, XLMiner partitions the data set before running the prediction method.

IV. NORMALIZATION TECHNIQUES

Normalization is the process of adjusting values measured on a different scale to a common scale. Normalization allows comparison of corresponding values of different data-sets. Without normalization, our data would be unscaled and hence highly intricate to calculate and compare with other parameters. There are many Normalization techniques – Feature scaling, Standardized moment, Coefficient variation, Studentized residual, Student's t-statistic, Standard Score.This system we are used two Normalization techniques in data mining

- Z-Score Normalization technique
- Min-Max Normalization technique

Z-Score Normalization technique: the Z-Score method or standard score method is presented, which normalizes each score to its number of standard deviations that it is distant from the mean score.

$x-\mu/\sigma$=x-mean/standared deviation

This formula rescales each of a feature's value in terms of how many standard deviations they fall above or below the mean value. The resulting value is called a zscore. The z-scores fall in an unbounded range of negative and positive numbers. Unlike the normalized values, they have no predefined minimum and maximum.

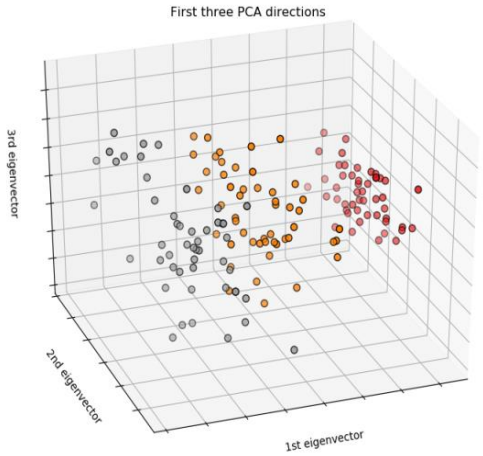
Min-Max Normalization technique: Min-Max is a technique that helps to normalize the data. It will scale the data between 0 to 1.

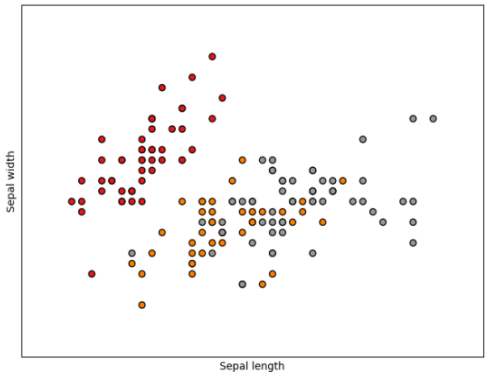
Xnew=x-min(x)/max(x)-min(x)

Normalized feature values can be interpreted as indicating how far, from 0 percent to 100 percent, the original value fell along the range between the original minimum and maximum.

V. IRIS DATASET

This data sets consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal length, stored in a 150x4 numpy.ndarray. The rows being the samples and the columns being: Sepal Length, Sepal Width, Petal Length and Petal Width. The below plot uses the first two features.





Data set:

	sepal_length	sepal_width	petal_length	petal_width	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3.0	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5.0	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5.0	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3.0	1.4	0.1	setosa
15	4.3	3.0	1.1	0.1	setosa
16	5.8	4.0	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa

VI. CONCLUSION

In this paper, we are used two normalization techniques that are Z-Score Normalization and Min-Max Normalization. Using these techniques computed the average prediction efficiency of K-nearest neighbor algorithm using the two normalization techniques and concluded the one technique with the highest efficiency.

REFERENCES

[1] Hui-Ling Chen, Bo Yang, Gang Wang, Jie Liu, Xin Xu, Su-Jing Wang and Da-You Liu

[2] Tam, Kar Yan, and Melody Y. Kiang. "Managerial applications of neural networks: the case of bank failure predictions. Radnaabazar Chinchuluun, Won Suk Lee, Jevin

[3] Bhorania and Panos M. Pardalos "Clustering and Classification Algorithms in Food and Agricultural Applications: A Survey"

[4] Reese, Heather, et al. "Applications using estimates of forest parameters derived from satellite and forest inventory data.

[5] "Ke-Wei Huang and Zhuolun Li- "A multi-label text classification algorithm for labeling risk factors in SEC form 10-K"

[6] Bijalwan, Vishwanath, et al. "KNN based machine learning approach for text and document mining. H. V. Jagadish Beng Chin Ooi, Kian-Lee Tan, Cui Yu, Rui Zhang

[7] Kyung-Chan Lee, Seung-Shik Kang, Kwang-Soo Hahn "A Term Weighting Approach for Text Categorization" 2005.

[8] Petra Perner "Advances in Data Mining: Applications and Theoretical Aspects:" 14th industrial conference , ICDM 2014 St Petersburg, Russia, 16-20, 2014 preceedings Tao Mei, Nicu Sebe, Shuicheng Yan, Richang Hong,

[9] Shipeng Li, Abdulmotaleb El Saddik, Meng Wang, athal Gurrin"Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images." 2013,2016.

[10] Mengfan Tang, Feiping Nie, Siripen Pongapaichet, Ramesh Jain "Semi-supervised learning on large-scale geotagged photos for situation recognition", 2017

[11] Zhaoqiang Xia, Jinye Peng, Xiaoyi Feng, Jianping Fan "Multiple Instance Learning for Automatic Image Annotation" 2013.

[12] Li,B, Lu,Q, Yu,S "An adaptive k-nearest neighbor text categorization strategy", 2004

[13] Li Baoli, Yu Shiwen, and Lu Qin "An Improved k-Nearest Neighbor Algorithm for Text Categorization", 2003

[14] Li Baoli, Chen Yuzhong, and Yu Shiwen, 2002. A Comparative Study on Automatic Categorization Methods for Chinese Search Engine [A]. In: Proceedings of the Eighth Joint International Computer Conference [C]. Hangzhou: Zhejiang University Press, 117-120.

[15] Yang Y. and Liu X., 1999. A Re-examination of Text Categorization Methods [A]. In Conference on Research and Development in Information Retrieval [C]. 42-49.