An Enhanced Performance of K-Nearest Neighbor (K-NN) Classifier to Meet New Big Data Necessities

Ihab L. Hussein Alsammak, Ministry of Education, Directorate General of Education of Karbala ehablaith@Gmail.com
Humam M. Abdul Sahib, Department of Electrical and Electronics Engineering, College of Engineering, University of Kerbala, Iraq humam.alkaabi@uokerbala.edu.iq
Wasan H. Itwee, Ministry of Construction Housing and Public Municipalities General Directorate of Sewerage/ Sewerage Karbala Directorate Wassan_iraq@yahoo.com

ABSTRACT:
The rapid increase in the growth of text information over the past two decades has led to the need for the use of text classification techniques, particularly in the area of information retrieval, data mining and data management. The precise results and simplicity of the K-Nearest Neighbor Classification Algorithm (K-NN) in knowledge mining is the reason that made it one of the most important classification algorithms used in many tasks such as pattern recognition, regression, and text classification. Through experiments and analysis of the results of the use of the traditional algorithm of the (K-NN), there are some deficiencies in their performance, especially when the data are large such as the algorithm was unable to process big data by rapid extraction with minimal storage space and generate useless samples computation and probability problems.

In this paper, we have developed an enhanced algorithm and get the best results and perform better than that in the traditional algorithm. The significant improvement in our model performance is due to the improvement by removing unnecessary computational samples in the traditional algorithm. The performance is further improved by using the lost value computational method to define results as a prelude to avoid wasting time by correcting and filtering noise, examining the database, and eliminating unwanted records. Additionally, the inverse logarithmic function was used to solve the probability problems the algorithm encounters. The experimental results showed the efficiency of the modified algorithm in reducing the sample size and speeding up the search for the required data.

Keywords: Data Mining, Classification, K-Nearest Neighbors, Noise Filtering, Missing Value, Logarithmic Function

1- Introduction and Motivation
The rapid increase in the growth of text information over the past two decades has led to the need for the use of text classification techniques, particularly in the area of information
retrieval, data mining and data management. Advanced methods for analysis and data mining (DM) are the most important tools used in all areas that used big data and in order to obtain accurate and relatively inexpensive information, especially for exploration and evaluation. In order to obtain accurate and relatively inexpensive information, the use of modern technology and the great progress in this area has helped to increase the amount of data and benefit from it and this helps companies in different domains of science [1]. Despite all of this technology, big storage difficulties are still one of the biggest hardness faced by large companies. However, researchers and companies are confronting difficulties in analyses these huge amounts of data. There are many classification methods used to classify the text, and the most well-known are: Bayesian, Decision Tree, Support Vector Machine (SVM), Linear Least Squares Estimator (LLSF), Neural Network and K Nearest Neighbor (KNN). Among these methods, the most precise results and simplicity of the K-Nearest Neighbor Classification Algorithm (K-NN) in knowledge mining is the reason that made it one of the most important classification algorithms used in many tasks such as pattern recognition, regression, and text classification [1].

One of the most important techniques used in various domains to extract data is the extraction of data (machine learning). Through this technology have been predicated totals for data examples [2]. Based on new technologies recently depend on a huge database, hidden patterns and relationships had been discovered [3]. When these technologies are applied to dense data there are failures in the areas of large and new data because of the complexity of the databases. Which in turn exceed the arithmetic, complexity and constitute an obstacle to the extraction of accurate data [4].

Due to the urgent need to convert the original data entered by users into data can be processed, in other words, are processed later to obtain valuable insights for this was used the term smart data and recently noticed the increasing need for these data, which was discovered in 2015 by Gartner, Inc. [5]. The reports Which were used by this data to be useful for the analysis of advanced and extract useful data only on the grounds that the databases contain all the data is of interest to companies or researchers or the last beneficiary from databases, which contribute to reduce the cost of data from the storage side in addition to the significant impact resulting in the handling of records in applications to ensure the success of data extraction [6]. Advanced analysis and data mining techniques are the most important tools used in all fields especially exploration and evaluation, which contribute to raising the means of development, updating and raising the efficiency of the final results and at the lowest cost[7] [8]. The record is based on information for those records, using one of the data mining
methods, such as K_nearest neighbor algorithm, to more accurately predict the results and identify the most important factors affecting the finally results[9].

Through experiments and analysis results used of the traditional algorithm for (K-NN), there are some deficiencies in their performance, especially when the data are large such as the algorithm was unable to process big data by rapid extraction with minimal storage space and generate useless samples computation and probability problems.

In this paper, we have developed an enhanced algorithm and get the best results and perform better than that in the traditional algorithm. The significant improvement in our model performance is due to the improvement by removing unnecessary computational samples in the traditional algorithm. The performance is further improved by using the lost value computational method to define results as a prelude to avoid wasting time by correcting and filtering noise, examining the database, and eliminating unwanted records. Additionally, the inverse logarithmic function was used to solve the probability problems the algorithm encounters.

This paper organized: Section 2 presents the previous related studies; in Section 3 motivation for a new approach, in Section 4 comparison K-NN algorithm with another classification algorithm, in Section 5 improve performance K-Nearest Neighbor, in section6 we provide the performance evaluation, in Section 7 we provide the experimental results, in Section 8 conclusions.

2- Literature Review

In this section we introduce an overview of important works in literature. To our knowledge, there are only a few works that explore separate acting-based classification strategies for a text classification task.

It is worth noting that some of the related works on the literature that interested in the texts classification of apply separate methods of feature engineering to extract effective features for use with the established classification methods. However, this has not been extended to classification methods.

With the rapid growth of the amount of databases, classification has become the main technology in the areas of information recovery, data mining and management. Now, in this aspect, multiple classification methods have been proposed, and among these methods, the K-NN algorithm method is considered private and effective as it was one of the most important algorithms in extracting data, which was popularly distributed in classification and used widely for two main reasons, the simplicity of the algorithm and the accurate results,
especially when dealing with a large database. k-NN algorithm classifier determines of checking data by explore at the k-NN in the data. The class of the plurality NN is then reported as the stamp of data [7].

Numerous previous research and studies concerned with databases proved the effectiveness of the K-NN algorithm to collect the required information easily, as opposed to its counterparts. Among these studies was Scott and Wade Wayne 1999 study, Iren Kerensky 2010 [10] where researchers relied on K-NN algorithm to collect information from large databases as quickly as possible. Another study aims to take advantage of the results of previous studies of the K-NN algorithm that were relied upon in extracting data to predict the phenomenon of colleges students dropping out [11].

The algorithm has been defined by some researchers such as Kaplan, Henlin (2010), Parkhouse and Tashiro (2010) and Wang, Chen and Liang (2011) as an existing relationship between a network of users and the methodology of this algorithm has been developed to play a major role in data mining and corporate earnings management [12].

The basis of the work of the K-NN algorithm was adopted and developed to discover new algorithms, as Dodani did, where he introduced the WKNN algorithm, this new algorithm was not satisfactory for the presence of extreme values [13]. This method reduces the weight of the K-NN algorithm compared to WKNN, thereby improving classification performance. Researcher Gou introduced new algorithm DWKNN depended on K-NN algorithm but DWKNN is also ineffective in dealing with asymmetric classes [14].

Researchers Zuo and Zhang have developed the work of the nearest neighbor algorithm to create a new algorithm called (KDF-KNN) to solve complex problems in databases [15]. In addition to what was previously mentioned there is a study a more in depth explanation of the basics of the K-NN algorithm was provided by Langley Pat as they were described in an indicative manner based on multiple databases [11].

Das researcher studied the pros and cons of the K-NN algorithm to present the BDSFS algorithm, which was considered a hybrid algorithm intended to reinforce decision-making decisions to choose distinct elements but this algorithm is a multi-category data problem interface and here it was not quite suitable for multi-category data sets.

3- Modified approach
One of the most popular and generally used techniques in DM is an algorithm K-Nearest Neighbor whose work depends on the concept of similarity. According to this mechanism, close data with similar patterns are grouped in a single row, and its work is like to the slow learning algorithm. This does not lead to special data on the training stage, because of the classification of invisible data and find them depending on the nearest ones. For this reason this algorithm suffers many disadvantages, especially when dealing with huge databases. In addition to this was pursued increase in the cost of accounting, noise and storage, noting the inability to handle enormous and incomplete information. The importance of this algorithm in the classification of data led to the researchers to processing the obstacles of this algorithm, get rid of the weaknesses and developed to obtain smart data (useful) of data, taking into account eliminates unnecessary or redundant data [17].

In this paper, we develop the traditional algorithm to improve its performance and be of use to researchers who deal with big data without defects and problems to ensure the final results. By studying the algorithm there are two main limitations in the algorithm: First, KNN algorithm using only probabilities to predict instance labels which outcome into undesirable performance on imbalanced data, the second limitation, the KNN algorithm uses all the training data, that mean the runtime is become slow. We will improve the extraction of data while reducing the storage space required, which is easy to be processed later. So, you must check the database and get rid of unwanted records before entering the accounts to avoid wasting time, and then use the inverted logarithm function to solve probability problems and remove extra data and irrelevant data. Now can explain K-NN classifier in following example, figure 1 shows that there are three classes: class A, class B and class C. Now, it is required to find out the class label for data sample P. In this example, a value of K=5. After calculating the Euclidean distance for each pair, four nearest neighbors were observed within the class A. and only one nearest neighbor were observed within the class C. So, the sample P is assigned to class A as it is the principal class for that sample [18]. The KNN algorithm is as follows: Training is very fast. You can learn complex target functions and do not lose information.
4- K-NN algorithm with another Classification algorithm

Through many studies, in aspects of classification methods such as Linear Least Squares Estimator (LLSF), Support Vector Machine (SVM), Neural Network, K Nearest Neighbor (KNN), Bayesian and Decision Tree. Considered KNN technology is a simple and efficient to deal with big database. Through several previous studies on the comparison between the KNN algorithm and the decision tree algorithm, being the most common among the classification algorithms, we can prepare a table that shows the difference between the KNN techniques and the decision tree. In Table 1, we can notice the difference between the two algorithms depending on using the WEKA tool table. Table 2 shows a summary of the results of accuracy.

**Table1: the comparison between K-NN algorithm and Decision Tree algorithm depended on Parameters**

| No | Parameter                  | K-NN                        | Decision Tree            |
|----|---------------------------|-----------------------------|--------------------------|
| 1  | Deterministic/Non-Deterministic | Non-deterministic           | Deterministic            |
| 2  | Effectiveness on Small data base | Large data base             |                         |
| 3  | Accuracy                  | Provides high accuracy.     | High accuracy            |
| 4  | Dataset                   | It can’t deal with noisy data. | It can deal with noisy data. |
| 5  | Speed                     | Slower for large data base. | Faster for all data      |
Table 2: summary the results of accuracy of classifiers

| No | dataset            | Size of Dataset       | KNN technique | Decision Tree technique |
|----|--------------------|-----------------------|---------------|-------------------------|
| 1  | Weather Nominal    | Small (14 instances)  | 100%          | 100%                    |
| 2  | Segment Challenge  | Medium (1500 instances) | 100%          | 99%                     |
| 3  | Supermarket        | Large (4627 instances) | 89.842%       | 63.713%                 |

5- Description of Improved Performance K-Nearest Neighbor

5.1- Problem statement

As is well known, the closest neighbor algorithm is considered a lazy learning method as it does not require an explicit training stage when dealing with big data. In previous studies, the DIRP method was used to measure the number of the nearest neighbors, and the distance to the Euclidean was based. By studying the algorithm it turns out there are two main flaws in the algorithm: KNN algorithm using only probabilities to predict instance labels which outcome into undesirable performance on imbalanced data, the second flaw the KNN algorithm uses all the training data that mean the runtime hence is become slowly.

1) Correction and noise filtering: after dealing with large data and tracking noise in the data, it was found that this problem is unavoidable. This in turn affects the efficiency of the data and hinders the process of data collection. These data are in the form of small groups of noise information in a certain area but originally belong to another category. To solve this problem, we can follow the method of disabling and removing layers from certain categories and increasing overlap between them. However, the final results of the extracted data may not be sufficiently strong to derive valuable knowledge.

The removal of the noisy measures after the identification of the training data is removed and indifferent is an effective and main step. In the specialized works, two changed types of noise were found: the first attribute noise and the second-class noise and these errors occur as a result of the introduction or naming of the sample in error, the second error to the rot in one or more features, but these problems have been dealt with in many ways useful and effective. Element noise rests to be less than discovered field.
By mentioning previously the K-NN classifier original is similarly affected through noisy data, the noise detection and isolation from the correct data and then removed based on the idea of distance-based relationship of the algorithm can illustrate the idea more easily can remove all training examples that do not match the class label based on the principle of search for the majority of the nearest neighbors. It has been proposed to change the names of the noisy groups to find the noise filters based on the algorithm, which is called a new label, namely edition-based models.

2) Missing values imputation: as mentioned previously about the problem of misinformed input data in data masses, many data in their KNN attribute values are lost in a simplified way, this data is lost. This error occurs as a result of errors in the equipment used or due to errors in the data entry itself.

To develop work, the extraction of data while reducing the storage space required for storage, which is easy to be processed later, and thus we will eliminate the data that are loud and the data is irrelevant.

5.2-The Proposed Modification

Depend on the K-NN algorithm, which one of its characteristics dealing with incomplete data where the problem of addressing the lack of data has been solved has been studied after distinguish the features of this algorithm two types of data that necessity to be processed the lacking data and noisy data. One of the maximums applied attribution approaches is created on the k-NN, after finding the incorrect data; each case should contain the incorrect data. New KNN is calculated based on most common values. Then the nominal values are selected based on all the neighbors for the numerical values after the data are extracted.

Negatively on the results obtained from mining through many studies and to search for the best solution found that the simplest way to solve this problem is to ignore the examples that contain them, but when dealing with the data affected by a large method is not practical. The procedure taken to get rid of false data is by calculating the wrong values by estimating them, but it should be noted that most of the time these features are linked to each other which does not work independently and this helps to study the qualities and determine the relations between these qualities to be rounded.

The first objective of reducing the data based on the chief database is to obtain accurate smart data without loss of data, through this process, we will improve the extraction of data while reducing the storage space required for storage, which is easy to be processed later, and thus we will eliminate the data that is irrelevant. Depending on the principle of the
perspective of attributes to reduce the data is typical selection and typical extraction, we
discovery instance shorthand methods at the instance level. This instance level is as a rule
divided into instance selection is limited to making a subset. Figure 2 shows the progress of
the proposed algorithm process, which shows the work steps as a flowchart.

![Figure 2: The Flowchart](image)

6- **Implement Modified K-NN algorithm**

By studying the algorithm as mentioned earlier, there were two main defects in the
algorithm. In the proposed algorithm the main objective is to solve these errors. Relying on
database checking and disposal of unwanted records before entering calculations to avoid
wasting time, and then using the inverted logarithm function to solve probability problems.
We will improve data extraction while reducing required storage space, which is easy to
process at a later time, and thus we will remove loud data and irrelevant data. In this study,
we used different sets of data, and the equations were reorganized in order to solve the test of
the learning ideas. Initially to see the process and performance improvement we impose a
small database and later to prove the final results in order to verify the correctness and
validity of the algorithm, was used 7500 news essays from 8 categories of Sina websites to
prove test the improved algorithm. 6,500 were used as practice samples and the residual
1,000 and the results are clear to the user with a follow-up that we can easily apply to large databases. The steps of proposed K-NN algorithm applied to the proposed dataset can be summarized based on the basic equation to solve the problem:

- Check the database in advance and delete all inaccurate records.
- Select the nearest neighbor number as K.
  \[ M(k) = \frac{1}{\log_2(1+k)} \]
- According to the apostolic distance between the explorer record and the nearest neighbor.
  \[ d(x_i, x_j) = \sqrt{\sum_{t=1}^{T} [X_{ti} - X_{tj}]^2} \]
- Rank the spaces by giving them the smallest distance to the highest distance.
- Select the nearest neighbors based on a minimum distance \( M(k) \) and k-th.
- Collect the category for nearest neighbors.
- Calculate the arithmetic mean of the nearest neighbor as a predictor value for the status of the explorer record.
- We continue to estimate the target function of the exploratory records.
- Calculate the RMSE value for the square root error rate for each K value if it is less than the previous station if not: Go to step 1.

Application Example: database and an objective test of two characteristics (Informative and Time Communion) as shown in the following table3.

|    | \textbf{X1= Informative} | \textbf{X2=Time Communion} | \textbf{Y=Classification} |
|----|--------------------------|-----------------------------|--------------------------|
| 1  | 7                        | 7                           | Bad                      |
| 2  | -1                       | 0                           | error                    |
| 3  | 3                        | 4                           | good                     |
| 4  | 1                        | 4                           | good                     |
| 5  | 7                        | 4                           | Bad                      |
| 6  | 0                        | -1                          | error                    |
| 7  | 6                        | 7                           | Bad                      |
| 8  | 6                        | 6                           | Bad                      |
The risk factors used in the prediction taken from the database records suggested. It was noted in the sample that a selected sample met certain conditions to meet the study objectives. And does not represent all recorded.

Now the being tested and for the purpose of obtaining a final result of the report apply the algorithm as follows:

- Determine the number of closest neighbors and assume that \( k = 1 \).
- Check the database in advance and delete all inaccurate records that mean the second record will be deleted.
- Calculate the distance between the query and all the training samples, since it is not counted
- distance we calculate the distance box that is faster for the calculation without using the root as shown in the table 4:

|   |   |   | error |
|---|---|---|-------|
|9  | -1| -1| error |
|10 | 4 | 3 | good  |
|11 | 4 | 1 | good  |
|12 | 7 | 6 | bad   |
|13 | 0 | 0 | error |
|14 | --| --| error |
|15 | 3 | 2 | good  |

The distance query to database:

| X1 | X2  | The distance box for the query |
|----|-----|-------------------------------|
| 7  | 7   | \((7-3)^2 + (7-7)^2 = 16\)    |
| 7  | 4   | \((7-3)^2 + (4-7)^2 = 25\)    |
| 3  | 4   | \((3-3)^2 + (4-7)^2 = 9\)     |
| 1  | 4   | \((1-3)^2 + (4-7)^2 = 13\)    |
| 6  | 7   | \((6-3)^2 + (7-7)^2 = 9\)     |
| 6  | 6   | \((6-3)^2 + (6-7)^2 = 10\)    |
| 4  | 3   | \((4-3)^2 + (3-7)^2 = 17\)    |
| 4  | 1   | \((4-3)^2 + (1-7)^2 = 37\)    |
| 7  | 6   | \((7-3)^2 + (6-7)^2 = 17\)    |
| 3  | 2   | \((3-3)^2 + (2-7)^2 = 25\)    |
- Arrange the distance to determine the nearest neighbors starting with the lowest distance and to up distance and depending on the K value, after that the inverted logarithmic function is using to solve the probabilities problems as shown in table 5.
- Now collect the closest category of neighbors taking into consideration that the second row of the nearest neighbors not within the classification of this sample is greater than k = 3 as shown in the table 6.

**Table 5: Arrange the Distance**

| X1 | X2 | The distance box for the query | The minimum distance rank | It is included within 3 closest neighbors? |
|----|----|--------------------------------|---------------------------|------------------------------------------|
| 7  | 7  | \((7-3)^2 + (7-7)^2 =16\)      | 3                         | Yes                                      |
| 7  | 4  | \((7-3)^2 + (4-7)^2 =25\)      | 4                         | No                                       |
| 3  | 4  | \((3-3)^2 + (4-7)^2 =9\)       | 1                         | Yes                                      |
| 1  | 4  | \((1-3)^2 + (4-7)^2 =13\)      | 2                         | Yes                                      |
| 6  | 7  | \((6-3)^2 + (7-7)^2 =9\)       | 1                         | Yes                                      |
| 6  | 6  | \((6-3)^2 + (6-7)^2 =10\)      | 1                         | Yes                                      |
| 4  | 3  | \((4-3)^2 + (3-7)^2 =17\)      | 3                         | Yes                                      |
| 4  | 1  | \((4-3)^2 + (1-7)^2 =37\)      | 4                         | No                                       |
| 7  | 6  | \((7-3)^2 + (6-7)^2 =17\)      | 3                         | Yes                                      |
| 3  | 2  | \((3-3)^2 + (2-7)^2 =25\)      | 4                         | No                                       |

- We take the largest group of the nearest neighbor class as a case predictive value 1 notes 2 good and 1 bad.

**Table 6: Closest Category of Neighbors**

| X1 | X2 | The distance box for the query | The minimum distance rank | It is included within 3 closest neighbors? | Category of the nearest neighbors |
|----|----|--------------------------------|---------------------------|------------------------------------------|----------------------------------|
| 7  | 7  | \((7-3)^2 + (7-7)^2 =16\)      | 3                         | Yes                                      | Bad                              |
| 7  | 4  | \((7-3)^2 + (4-7)^2 =25\)      | 4                         | No                                       | ------                           |
| 3  | 4  | \((3-3)^2 + (4-7)^2 =9\)       | 1                         | Yes                                      | Good                             |
| 1  | 4  | \((1-3)^2 + (4-7)^2 =13\)      | 2                         | Yes                                      | Good                             |
| 6  | 7  | \((6-3)^2 + (7-7)^2 =9\)       | 1                         | Yes                                      | Bad                              |
| 6  | 6  | \((6-3)^2 + (6-7)^2 =10\)      | 1                         | Yes                                      | Bad                              |
7-Evaluation of New-KNN algorithm

After making the required improvements to the KNN algorithm to eliminate the problem of undesirable performance and get rid of unwanted records before entering the calculations to avoid wasting time, and then use the inverted logarithmic function to solve probability problems. Through previous research adopted to find out the time used to implement the algorithm compared to the time of the optimized algorithm to show the obvious difference as shown in Figure 3, a value of K is suggest (1, 3, 5, and 10). The database is indicated by the symbol (14n, 1200n, 1500n, 2400n, 4800n) as shown in figure 4.

|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
|4  | 3 |  \((4-3)^2 + (3-7)^2 = 17\) | 3 | Yes | Good |
|4  | 1 |  \((4-3)^2 + (1-7)^2 = 37\) | 4 | No | Good |
|7  | 6 |  \((7-3)^2 + (6-7)^2 = 17\) | 3 | Yes | Bad  |
|3  | 2 |  \((3-3)^2 + (2-7)^2 = 25\) | 4 | No | Good |

To evaluate the New KNN algorithm, we used 5 data sets from UCI machine learning repository. These data sets have been selection due to their sway in the propriety. The data sets from UCI are: Glass data set was given by USA, Mushroom this dataset was given by Audubon Society Field Guide, Bupa dataset was donated by BUPA Medical Research Ltd, E.Coli this dataset contains protein localization sites and Pima dataset was given by National Institute of Diabetes and Digestive and Kidney diseases. The finally results (accuracy in %) as show in figure 4 that our algorithm New-KNN produces better results as compared to the traditional KNN.
Figure 4: Accuracy of Proposed Knn

8- Conclusions

From previous studies, the K-NN algorithm can be considered one of the ideal ways to find smart data from big databases. However, as we have pointed out, there are some drawbacks to this algorithm to get accurate and useful final results. These points were discussed to extract valuable smart data. Significantly, we need to define strategies for setting the K-NN to correct the data set that lacks lost or noisy values, and how to reduce the data that contributes effectively to the speed of implementation and reduced implementation time. In this paper, we have developed the work of the K-Nearest Neighbor algorithm by relying on a valid, easily manageable data set to reduce the time needed to obtain the final results, so we can consider it a cost-effective and efficient way to easily collect the required information. To evaluate the work of the proposed KNN algorithm, the final results showed that the proposed algorithm gave better accurate results compared to the traditional KNN algorithm at an ideal time.

REFERENCES

1. The Playground. (May 2016). http://blog.kaggle.com/2013/09/25/the-playground.
2. Francis R Bach, Gert RG Lanckriet, and Michael I Jordan.. Multiple kernel learning, conic duality, and the SMO algorithm In Proceedings of the twenty-first international conference on Machine learning. ACM, 6,2004.
3. Tianqi Chen and Carlos Guestrin.. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 785–794, 2016.
4. V. Marx, “Biology: The big challenges of big data,” Nature, vol. 498, no. 7453, pp. 255–260, Jun. 2013.
5. K. Sun, H. Kang, and H.-H. Park, “Tagging and classifying facial images in cloud environments based on KNN using mapreduce,” Optik International Journal for Light and Electron Optics, vol. 126, no. 21, pp. 3227 – 3233, 2015.

6. Sriram Kankatala E-mail(s): kasr13@student.bth.se, Performance Analysis of KNN on large datasets using CUDA & Pthreads. Master Thesis Electrical Engineering September 2015.

7. Contemporary Issues In Education Research – Second Quarter 2014 Volume 7, Number 2 Copyright by author(s); CC-BY 129 The Clute Institute The Impact Of Social Networking: A Study Of The Influence Of Smartphones On College Students Abir S. Al-Harrasi, Sultan Qaboos University, Oman Ali H. Al-Badi, Sultan Qaboos University, Oman.

8. J. Maillo, S. Ramirez, I. Triguero, and F. Herrera, “kNN-IS: An iterative spark-based design of the k-nearest neighbors classifier for big data,” Knowledge-Based Systems., vol. in press, 2016.

9. Iraqi Journal of Statistical Sciences (18)2010,[236-209]

10. J. Dean and S. Ghemawat, “Map reduce: A flexible data processing tool,” Communications of the ACM, vol. 53, no. 1, pp. 72–77, 2010

11. F. Iafrate, “A journey from big data to smart data,” in Digital Enterprise Design & Management: Proceedings of the Second International Conference on Digital Enterprise Design and Management DED&M, pp. 25–33, 2014, Cham, 2014.

12. Zuo, Wangmeng, David Zhang, and Kuanquan Wang. "On kernel difference-weighted k-nearest neighbor classification." Pattern Selection of relevant features in machine learning, Proceedings AAAI Fall Symposium on Relevance, New Orleans, LA (1994), pp.140–144.

13. Dudani, Sahibsingh A. "The distance-weighted k-nearest-neighbor rule." Systems, Man and Cybernetics, IEEE Transactions on 4(1976): 325-327.

14. Gou, J., et al. "A new distance-weighted k-nearest neighbor classifier." J. Inf. Comput. Sci 9 (2012): 1429-1436.

15. C. Zhang, F. Li, and J. Jestes, “Efficient parallel knn joins for large data in mapreduce,” in Proceedings of the 15th International Conference on Extending Database Technology, New York, NY, USA: ACM, pp. 38–49, 2012.

16. C. P. Chen and C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on big data,” Information Sciences, vol. 275, pp. 314 – 347, 2014.

17. Sriram Kankatala E-mail(s): kasr13@student.bth.se, Performance Analysis of KNN on large datasets using CUDA & Pthreads. Master Thesis Electrical Engineering September 2015.

18. Chakraborty, Dipanjan(2008), K-Nearest Neighbor Learning.