A New Hybrid KNN Classification Approach based on Particle Swarm Optimization

Reem Kadry[1]
Faculty of Computer Science
October University of Modern Science and Arts, Egypt

Osama Ismael[2]
Faculty of Computers and Artificial Intelligence
Cairo University, Egypt

***Abstract*—K-Nearest Neighbour algorithm is widely used as a classification technique due to its simplicity to be applied on different types of data. The presence of multidimensional and outliers data have a great effect on the accuracy of the K-Nearest Neighbour algorithm. In this paper, a new hybrid approach called Particle Optimized Scored K-Nearest Neighbour was proposed in order to improve the performance of K-Nearest Neighbour. The new approach is implemented in two phases; the first phase help to solve the multidimensional data by making feature selection using Particle Swarm Optimization algorithm, the second phase help to solve the presence of outliers by taking the result of the first phase and apply on it a new proposed scored K-Nearest Neighbour technique. This approach was applied on Soybean dataset, using 10 fold cross validation. The experiment results shows that the proposed approach achieves better results than the K-Nearest Neighbour algorithm and it's modified.**

Keywords*—*K-nearest neighbour; outlier; multidimensional; particle swarm optimization; scored k-nearest neighbour

I. INTRODUCTION

The rapid development in data collection techniques and storage technologies enables organizations to reserve large amount of data. With the help of data mining algorithms the quality of decision making may be supported and human error is avoided. One of the most popular algorithms of data mining is K Nearest Neighbour (KNN). KNN is a simple effective supervised classification algorithm that is easily to be implemented on different kinds of data. Also, KNN is called lazy learning algorithm while rest of classification algorithms are called eager ones. Precisely speaking, there is no explicit training phase in KNN. It starts working only when it gets an unseen tuple for classification [1].

Multidimensional data and outliers are two of the main problems in KNN algorithm. KNN algorithm does not work accurately with multidimensional data as the distance in each dimension is hardly calculated. Feature reduction and optimization of multidimensional data will help to increase the effectiveness of KNN. Moreover, the presence of outliers in the data to be mined by KNN affects the accuracy of the result. Outlier reduction in KNN is a common challenge in many researches in order to reduce the unwanted data set which is not relevant to the pattern.

Motivated by these facts, in the current work, an effective hybrid approach for pattern detection is devised which uses the output of Particle Swarm Optimization as an input to proposed new scored KNN algorithm. We make two modifications on PSO, the first modification is proposing a new way to calculate the learning rate that was inspired from the neural networks, the second one is a threshold accuracy percentage as a termination condition. We utilize the PSO to solve the multidimensional data and then using this optimized data as initialization to our new proposed scored KNN algorithm. The idea of the proposed scored KNN technique is near to weighted KNN, taking into consideration the calculation of mean and maximum distance for each k-nearest neighbor, then give score for every distance. The aim is to be close to the mean distance. This overcomes the outliers problem in KNN, because KNN equalize all neighbours regardless different distances. So that it will be helpful for achieving maximum accuracy in minimum computation time.

In order to employ formal and systematic evaluation of the work in this research, a set of experiments has been designed on Soybean data sets [18] using 10 fold cross validation.

II. RELATED WORK

KNN is one of the algorithms which performs classification regardless a past knowledge of the data set to be classified. Classification is done based on similarity. Despite the popularity and simplicity of KNN, it has some shortcomings affecting its performance. Many researches are done to enhance KNN performance, within multiple improvement directions; outliers reduction in [2][3][4], appropriate K value selection [5], integration of multiple algorithms [6] [7] [8] and more other improvements [9] [10][11].

Divya and Senthil [2] introduced a new outlier reduction approach based on distance learning for categorical attributes and introduced distance learning framework. The main problem in the introduced approach was the need to remove redundant or irrelevant features to improve classification performance and decrease cost of classification. So, Selahaddin and Ahmet [3] introduced a new density weighted KNN to reduce the effect of irrelevant data. They obtained the coefficient of density of each element of the training data set. Then, determined the relation of each test element based on the total of density coefficients of neighbour that belong to the same class. The problem of this approach was the amount of resources needed to run it, and so it needs more time and memory requirements. Also, Guo-Feng et al [4] proposed a new short term load forecasting mode based on weighted K nearest neighbour algorithm to achieve more satisfied accuracy. The problem with approach is the need to use an optimization technique in order to improve the forecasting accuracy.

Another enhancement of KNN algorithm is introduced by Natalia et al [5] to select adaptive number of nearest neighbour; for each test point in N, the method looks from 1 to M nearest neighbour at the same time, and finds the value k. This algorithm was applied on diabetes data and it is needed to be generally applicable to classification problem of different data sets to ensure its performance.

More enhancements were done by integration of multiple algorithms, Bahramian and Nikravanshalmani [6] proposed a new classification algorithm based on feature selection with genetic algorithm and combination of k nearest neighbour and Adaboost (a practical boosting algorithm works on classification problems in order to change a group of weak classifiers into a strong one) algorithms to increase the accuracy of classifying diabetes dataset. Also Chen and Hao [7] proposed a hybrid framework of weighted support vector machine and weighted KNN. First, establish detailed feature weighted SVM for data classification giving different weights for different features. Then, estimate the importance of each feature to get weights, this by computing the information gain. Finally, weighted KNN by computing k weighted nearest neighbours from the historical dataset. Reyhaneh Sadat Moayeri et al [8] proposed a hybrid predictive model to evaluate dental implants success using SVM, Neural Networks and KNN. The combined classifier aimed to be achieved a higher accuracy than using one classification algorithm, but the main issue that the data used was only for the medical field with specific attributes.

More enhancements done by Li Yu et al [9], they discussed the effect of distance function on KNN performance. They used four different functions; Euclidean, Cosine, Chi square and Minkowsky and compared the performance results on different medical datasets. Also, Yanpeng et al [10] introduced a multifunction nearest neighbour approach, by combing fuzzy similarity relations and class memberships. This approach gives an adaptively to deal with KNN classification problems.

Shubham Pandey et al [11], proposed a new technique called Modified K Nearest Neighbour (MKNN) which was based on assigning the class label of the required instance into K validated training data points. Then, compute the validation of the data sample in the train set. After that, perform a weighted KNN on any test sample. When they implemented their experiment on Soybean dataset, the reached average accuracy of the MKNN was 85.56 %. The issue was that they compared their results with the classic KNN only. In our research we will compare the results of the proposed approach; Particle Optimized K Nearest Neighbour, with the classic KNN and also with this Modified KNN.

Our proposed approach focuses on solving two main problems affecting the accuracy of KNN; which are multidimensional data and outlier reduction. This was achieved through the integration between Particle Swarm Optimization technique and the new Scored KNN approach, in order to introduce an effective hybrid approach for pattern detection called Particle Optimized Scored KNN (POSKNN).

III. Particle Optimized Scored KNN Approach (POSKNN)

KNN is considered to be simple, effective, intuitive and competitive classification algorithm in several domains. Despite KNN advantages, it has some limitations that can affect its performance. It is very sensitive to irrelevant or redundant features because all features contribute to the similarity and thus to the classification. By careful feature selection or feature weighting, this can be avoided [12]. Also, the presence of outliers will have an effect on the accuracy of classification. As the distance between an object and its neighbour increases, the more it will be considered as an outlier. Outlier reduction is considered as a common problem in KNN algorithm as it affects its accuracy.

Motivated by the above mentioned problems, we propose a novel hybrid approach called Partial Optimized Scored KNN. To build and utilize the model, two phases are undergone; first multi-dimensional feature selection phase. In which, we aim to select the most important features, the result of this phase is a reduced set of features. Second, outlier reduction phase, the input of it is the set of selected feature from phase one, and introducing a new scored KNN technique, aiming to give a score for each distance in order to determine the extreme distance. The next two subsections we will give a detailed description for these two phases.

A. *Mulit-Dimensional Feature Reduction Phase*

Feature reduction is a critical step in data pre-processing and important research content in data mining tasks such as classification. Feature selection is to effectively reduce feature dimension and improve classification accuracy and efficiency by deleting irrelevant and redundant features in data sets [13]. KNN algorithm has some problems to deal with multi-dimensional datasets. As the number of dimensions increases, the calculated distances are less considerable, so the performance of KNN results decreases.

The main goal of the multi-dimensional feature selection phase is to get rid of the redundant and irrelevant data. This will result in improving the classifier effectiveness and the classifier accuracy through increasing the true positive predictive values. In this research, multi-dimensional feature reduction phase is done through utilizing the Particle Swarm Optimization (PSO) technique. PSO has powerful convergence ability to the optimization value and it can be hybridized with other algorithms easily. PSO is a natural optimization technique based on the synchronization of the movement mechanism of swarms. PSO algorithm simulates animal’s social behavior including insects, herd, birds and fishes. These swarms conforms a cooperative way to find food, and each member in the swarms keeps changing the search pattern according to the learning experiences of its own and other members [14].

PSO is a computational searching algorithm which optimizes a problem by trying to improve different solutions on the basis of a specific quality measure. It solves a problem by having a population of different solutions (called particles). Each particle moves around in the search area according to a simple mathematical equation, with specific position and velocity for each particle. In each move a particle is affected by its local position and also is affected by the updated better position founded by other particles. The process is repeated till a stopping condition is founded which is the number of iterations. At each iteration of PSO, position and velocity of every particle is updated according to (1) and (2):

$$v_i(t+1) = v_i + C_1 \times rand_1 \times (pBest_i - x_i) + C_2 \times rand_2 \times (gBest - x_i) \quad (1)$$

$$x_i(t+1) = x_i + v_i(t+1) \quad (2)$$

Where, the initial position and velocity of particles are generated randomly within the search space. Rand_1 and rand_2 are generated random numbers ϵ [0, 1]. C_1 is the learning rate of personal experience and C_2 is the learning rate of global experience.

Some basic parameters may affect PSO performance, as the number of particles, number of iterations and learning rate. The number of particles used in PSO ranges from 10 to 100. There is no exact rule in literature for selection of swarm size. But normally, when the dimension of problem at hand increases, the swarm size should also be increased [15]. Too few particles prompt the algorithm to get trapped in local optima, while too many particles slow down the algorithm. Also, the number of particles has an effect on the computational complexity; if its value increase the time consumed to reach good optimization results increases. Based on the done experiments– as illustrated in the next section- we set the number of particles to 20.

The number of iterations in PSO represents the stopping condition of the algorithm. The termination condition may be maximum number of iterations, or termination when finding an acceptable solution to a given problem. We proposed a termination condition depending on the achieved accuracy, in order to decrease the time complexity. Several test cases were implemented in order to illustrate the effectiveness of the proposed termination condition. The algorithm is supposed to terminate when reaching threshold minimum accuracy equals to 94 %. In our case based on the proposed termination condition, the algorithm was terminated after a range from 32 to 37 iterations.

Also, the learning factor represented by the coefficients C1 and C2. To adjust the weight of empirical information of each particle, the coefficient C1 is used, and to adjust the weight of integrated information of the optimal particle in the current population, the coefficient C2 is used. If the value of C1 is too large, it will be easy to fall into local convergence. Also, if the value of C2 is too large, the algorithm will cost an expensive computation overhead and easily fall into iterative divergence. Conventional PSO algorithms usually sets C1 = C2 = 2 or other fixed constants [16]. In our research, we proposed a new way to calculate the learning rate inspired from the calculation of learning rate in neural networks, which depending on the number of iterations as illustrated in (3).

$$Learning\ rate = \frac{Number\ of\ particles}{Number\ of\ iterations} \quad (3)$$

As the number of iterations increases the learning factor decreases for every particle, that's why we divided by the number of iterations. When applying PSO to the used dataset with the proposed termination condition and the proposed learning rate, it gave better performance than applying it with constant values for iteration number and learning rate.

B. KNN Outliers Reduction Phase

Outlier reduction is considered to be an important problem in the field of data mining. In general, outlier reduction is the concept of searching for instances in a dataset which are inconsistent with the remainder of that dataset [17]. KNN accuracy is greatly affected by the presence of outliers. In our research, we aim to identify the unusual records in the used dataset by introducing the new scored KNN approach.

In the second phase; KNN Outliers Reduction Phase, we took the output of the optimized selected features from phase 1, as an input. First, we applied KNN algorithm by initializing the value of k equals to 2 then adjusting the value of k till k equals to 20. And we chose the value of k with respect to experienced test results to get optimized classification results. Then, calculate the distance using standard Euclidean distance. The result of this step will detect the outliers, as the distance to its k^{th} nearest neighbour is considered as the outlying score. Secondly, to solve the problem of outliers, new scored KNN is proposed by sorting the calculated distance in descending order. Thirdly, the mean distance is calculated and the maximum distance is obtained, to be used in the new scored KNN function as follows in (4):

$$New\ Scored\ KNN = 1 - \frac{|\,distance - mean|}{|\,maximum - mean|} \quad (4)$$

Where:

- Distance is calculated by the Euclidean distance: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}\ ^2$
- Mean: is the mean distance based on **K** parameter.
- Maximum: is the maximum distance based on **K** parameter.

Finally, get the sum of the scored distance for each class and the class with maximum score will be the predicted class. The main aim of introducing this function is to be close to the mean distance in order to give the outliers low score.

Our novel hybrid approach; Particle Optimized Score K-Nearest Neighbour is the result of combining 1st phase and 2nd phase. The experiment of the proposed approach will be discussed in details in the experiment section.

IV. Experimental Evaluation

In our experiments the new proposed scored KNN algorithm was run based on the output of applying the particle swarm optimization algorithm and compared the experiment results with the traditional and MKNN algorithm [11]. The used datasets is: "Soybean large dataset"; it is composed of 36 attributes, 15 classes and 684 instances [18].

All the experiments are executed using Matlab 2014. We applied Particle Swarm Optimization on Soybean dataset with the following inputs; number of particles N, number of iteration T and the acceleration coefficients c1 and c2. After making multiple experiments with different range of number of particles, we found that the best accuracy reached with N = 20. T is the maximum number of iterations, but we proposed using threshold accuracy equals to 94%, which can be achieved after around 30 iterations. But it was concluded that when the value of k is small the threshold accuracy could not be reached, so the algorithm takes more iterations and terminate with different accuracy values less than 94 %. Regarding the acceleration coefficients c1 and c2, also a new method for calculation was proposed, in which c1 and c2 will be equal to the number of particles divided by the number of iterations and it is updated every iteration.

PSO's parameters have a great effect to the convergence rate and the time for reaching the optimal solution. Fig. 1 shows the convergence curve of PSO algorithm, with x-axis representing the number of iterations and with y-axis representing the fitness value that minimize our optimization function (minimize number of features).

After the 1st iteration, the best fitness value of 0.383, was reached by the 7th particle. After the 5th iteration, the other particles were guided by the 7th particle to better best position, and achieved other best fitness value of 0.0315. PSO continued its iterations till reaching the threshold accuracy of 94 %, and converged to the best optimal solution at iteration number 30 with minimum fitness value of 0.243.

Fig. 2 shows that when reaching accuracy of 74 % the number of features were 29 with k value equals to 3, then when the accuracy was 80 %, the number of features was 26 with k value equals to 5. Then the number of features decreased gradually till 18 features, and reached the threshold accuracy of 94 % when k value equals to 14.

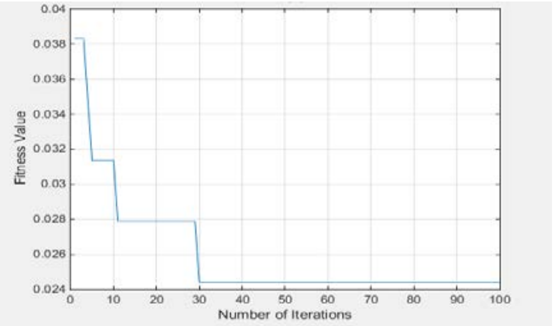

Fig. 1. PSO Convergence Curve.

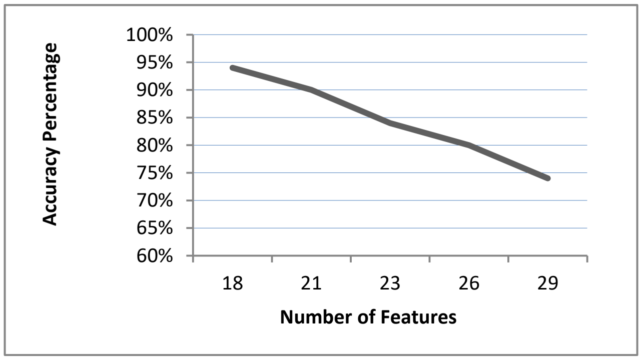

Fig. 2. Number of Iterations based on Percentage.

V. Evaluating Classifier Performance

An evaluation of the classifier performance was needed to know how accurate the classifier is, to predict the class label of tuples. In our experiments the following evaluation measures were used:

- The accuracy of a classifier.
- Error rate.
- Sensitivity and specificity.
- Precision and recall.

The percentage of test set tuples that are correctly classified by the classifier, shows the accuracy of a classifier on a given test set. Also, the error rate or misclassification rate of a classifier is simply 1 – accuracy. When the main class of interest is rare (e.g. in fraud reduction application), the class of interest (or positive class) is "fraud", which occurs much less frequently than the negative "non-fraudulent" class. In medical, there may be a rare class, such as "cancer". Therefore, other measures are needed, that recognize how the classifier can correctly predict the positive tuples (cancer = yes)) and how it can correctly recognize the negative tuples (cancer = no). To apply this, the sensitivity and specificity measures can be used sequentially. The true positive rate is defined as Sensitivity (i.e., the proportion of positive tuples that are correctly identified), while the true negative rate is defined as specificity (i.e., the proportion of negative tuples that are correctly identified). Finally, Precision is a measure of exactness, whereas recall is a measure of completeness. The Recall is also most likely known as sensitivity. These measures are defined as:

$$Accuracy = \frac{TP + TN}{P + N} \qquad Error\ rate = \frac{FP + FN}{P + N}$$

$$Sensitivity = \frac{TP}{P} \qquad Specificity = \frac{TN}{N}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

A comparison between the performance evaluation measures of the classical KNN with k values ranges from 3 to 14, MKNN and the new proposed POSKNN approach were conducted through a different set of experiments on the Soybean dataset. The experiments were applied to a different number of nearest neighbors (K) to verify the effect of varying number of K on the experimental results. Table I shows the average of evaluation of classical KNN, MKNN and POSKNN algorithm.

As shown in Table I, when applying POSKNN with different values of k equals 3, 7, 9 and 14 it gave highest average accuracy which shows how our classification is close to real value. It was concluded that when the k value is small the classifier could not reach the threshold accuracy of 94%, so it continued until the 100 iteration and then terminated with accuracy values less than 94%, and when the k value reaches 14, it terminated when reaching the threshold accuracy. POSKNN gave lowest error rate; this shows that the proposed approach gives minimum misclassification rate. Also, it gave highest recall, indicating that the number of correct results of classification is high. Moreover, POSKNN gave highest specificity showing that the number of actual negatives that is correctly classified is high.

Fig. 3 shows the accuracy evaluation of classical KNN, MKNN and POSKNN algorithm. When K was set to 3, the traditional KNN algorithm performs better than MKNN and POSKNN algorithm. As K increases PSOKNN becomes better than the KNN, and MKNN algorithms.

Based on the empirical studies, after applying different K values of 3, 7, 9 and 14, POSKNN reached an accuracy of 80.86 %, 91.8 %, 92.15 % and 94.21% respectively. When setting K to 14 the best accuracy was reached. When K is too small, the POSKNN classifier may be misleading because of noise in the data. On the other hand, when k is increased more than 14, the POSKNN classifier may misclassify the test instance because its list of nearest neighbours may include data points that are located far away from its neighbourhood, which may include points from other classes. On the other hand, traditional KNN and MKNN reach the best accuracy when K is set to 9, and as the value of K increases, the accuracy decreases.

Fig. 4 shows the average accuracy of KNN, MKNN and POSKNN with K value ranges from 3 to 14. It is observed that our proposed approach (POSKNN) outperform the other algorithms.

From the experimental results, it is noticed that the PSOKNN outperforms traditional KNN and MKNN in all cases, except when K is extremely small, the classifier may be misleading because of noise in the data.

TABLE I. AVERAGE PERFORMANCE EVALUATION MEASURES

Point of Comparison	KNN	MKNN	POSKNN
Accuracy	83.94 %	85.56 %	88.59 %
Error Rate	0.0662	0.0578	0.04854
Recall	75.48 %	77.5 %	79.98 %
Precision	88.4 %	89.3 %	91.2 %
Specificity	89.23 %	90.1 %	91.5 %
Sensitivity	75.48 %	77.5 %	79.98 %

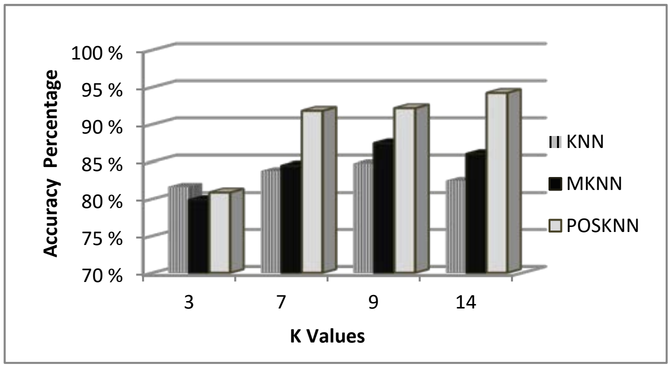

Fig. 3. The Accuracy Evaluation of Classical KNN, MKNN and POSKNN Algorithm.

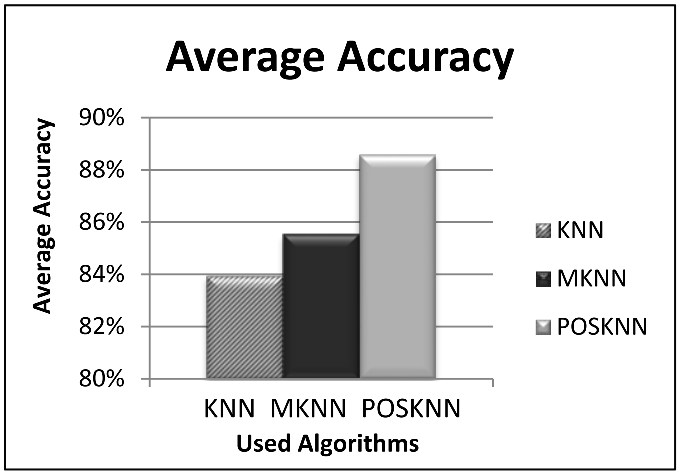

Fig. 4. Average Accuracy of the Three Algorithms.

Hypothesis testing was also used to ensure with a predefined level if the proposed approach has higher performance than the others [19]. T-test was used to check if the differences between the averages of the resulting values for KNN, MKNN and PSOKNN are statistically significant [20]. The results of carrying out the T-test, with using a confidence level equal 95% are illustrated in Table II. GraphPad software (http://www.graphpad.com/) was used for calculating T-test.

TABLE II. T-TEST RESULTS

KNN (average)	MKNN (average)	PSOKNN (average)	K Value	T Value	Two-tailed P value	Statistical significant
81.62	79.87	80.86	3	0.7020	0.5554	Not Statistically Significant
83.75	84.46	91.8	7	0.7788	0.5176	Statistically Significant
82.45	90.08	94.21	14	6.2054	0.0250	Extremely statistically significant.

T-test results demonstrate that when values of k were large (k>=7), the results were statistically significant, while for small values of k (k=3), the results were not statistically significant. When K is set to 14, the two-tailed P value is 0.0250. Therefore, the difference between the means is extremely statistically significant. In the case of K equals 3, the two-tailed P value is 0.5554, it states that the difference between means is not statistically significant. Finally, when K is set to 7, the two tailed P value is 0.5176; thus the difference between means is statistically significant.

VI. CONCLUSION AND FUTURE WORK

This paper proposes a new hybrid approach to improve the performance of KNN classifier, which is called Particle Optimized Scored K-Nearest Neighbour. The proposed approach applies Particle Swarm Optimization to solve the problem of multidimensional data in KNN. PSO is applied with two modifications; first, introducing threshold accuracy as a termination condition for the number of iterations, second, proposed a new technique to calculate the learning factor inspired from neural networks and customized on PSO algorithm.

In order to get the best performance, we applied a new Scored KNN on the optimized output of PSO. The new Scored KNN takes into account the value of the mean distance and this helped to solve the problem of outliers' problem in KNN. The new proposed hybrid approach was evaluated on Soybean dataset. Results showed that the proposed approach; Particle Optimized Scored K Nearest Neighbour (POSKNN), gave better accuracy than the classic KNN and the MKNN.

As a future work, an implementation of the proposed algorithm will be done on different dataset we need to investigate other optimization algorithms to be applied with the new proposed Scored KNN on the same dataset and compare results. Furthermore, we need to see the effect of applying the same proposed system using different distance measure such as Manhattan, Minkowski and cityblock. Moreover, we will calculate the execution time of the 3 methodologies; classical KNN, MKNN and POSKK, to ensure that the accuracy was increased without increasing the execution time.

References

[1] A. Lamba and D. Kumar "Survey on KNN and its variants", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, Issue 5, May 2016.

[2] K. Divya and N. Kumaran "Improved outlier detection using classic KNN algorithm", International Research Journal of Engineering and Technology, vol. 3, Issue 12, December 2016.

[3] S.Akben and A. Alkan "Density weighted K-nearest neighbors algorithm for outliers in the training set are so close to the test element", Journal of Electrical Engineering, vol. 3, Issue 3, pp. 150 – 161, 2015.

[4] G. Fan, Y. Guo, J. Zheng and W. Hong "Application of weighted K nearest neighbour algorithm for short term load forecasting ", Energies Journal, vol. 5, Issure 12, pp. 916, March 2019.

[5] N. Labuda, J. Seeliger., T. Gedrande and K. Kozak, "Selecting adaptive number of nearest neighbors in k-nearest neighbor classifier apply diabetes data", Journal of Mathematics and statistical Science, Issue 1, 2017.

[6] S. Bahramian and A. Nikravanshalmani, "Hybrid algorithm based on K-nearest algorithm and Adaboost with selection of feature by genetic algorithms for the diagnosis of diabetes", International Journal of Mechatronics, Electrical and Computer Technology (IJMEC), pp. 2977-2986, 2016.

[7] Y. Chen and Y. Hao, "A feature weighted support vector machine and K-nearest neighbour algorithm for stock market indices prediction", Expert System Applications, vol. 80, pp. 340–355, March 2017.

[8] R. Sadat, M. Khalili and M. Nazari "A hybrid method to predict success of dental implants", International Journal of Advanced Computer Science and Applizations, vol. 7, Issue 5, May 2016.

[9] L. Hu, M. Huang,, S. Ke and C. Tsai "The distance function effect on k‑nearest neighbor classification for medical datasets", SpringerPlus, vol. 1, Issue 5, August 2016.

[10] Y. Qu, C. Shang, N. Parthalain, W. Wu. and Q. Shen "Multi- functional nearest-neighbour classification", Soft Computing Journal, vol. 22, Issue 8, pp. 2717 – 2730, March 2017.

[11] S. Pandey, V. Sharma and G. Agrawal,"Modification of KNN algorithm", International Journal of Engineering and Computer Science, vol.8, Issue 11, pp. 24869 – 24877, November 2019.

[12] S. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) approach for predicting economic events: Theoretical background", International journal of engineering research and application, vol. 3, Issue 5, pp. 605 – 610, January 2013.

[13] J. Wang, J, Xu, C. Zhao, Y. Peng and H. Wang, "An ensemble feature selection method for high- dimensional data based on sort aggregation", System science & control engineering journal, vol. 7, Issue 2, pp. 32 – 39, May 2019.

[14] D. Wang, D. Tan and L. Liu, "Particle swarm optimization algorithm: an overview", Soft Computing, vol. 22, pp. 387 – 408, January 2018.

[15] A. Jorhedi and J. Jasni, "Parameter selection in particle swarm optimization: a survey", Journal of experimental & theoretical artificial intelligence, vol. 25, Issue 4, pp. 527-542, June 2013.

[16] M. Ren, X. Huang, X. Zhu and L. Shao, "Optimized PSO algorithm based on the simplicial algorithm of fixed point theory", Applied intelligence, vol. 50, Issue 7, pp. 2009 – 2024, July 2020.

[17] F, Benjelloun, A. Ousso, A. Bennani., S. Belfkih and A. Lachen, A., "Improving outliers detection in data streams using LiCS and voting", Journal of King Saud University- Computer and Information Sciences, pp. 1319 – 1578 August 2019.

[18] D. Grant, Nelson, R. Nelson, S. Cannon and R. Shoemaker, "SoyBase, the USDA-ARS soybean genetics and genomics database", Nucleic Acids Research, vol. 38, Issue suppl_1, pp. D843 – D846, 2009.

[19] T. Kim, "T test as a parametric statistic", Korean Journal of Anesthesiology, vol. 68, Issue 6, December 2015.

[20] F. Emmert and M. Dehmer "Understanding statistical hypothesis Testing: the logic of statistical inference", Machine learning & knowledge extraction, vol. 1, pp. 945-961, August 2019.