International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Predictive Accuracy of Modified Subtractive Clustering Algorithm on Large Dataset

Nidhi Pandey
Computer Science.
SRCEM, Banmore, (M.P.)
nidhipandeynidhi@gmail.com

Nirupama Tiwari
Asst. Prof. Computer Science.
SRCEM, Banmore, (M.P.)
girishniru@gmail.com

Abstract **– Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than amongst groups. This work reviews most representative subtractive clustering techniques on Matlab platform. The proposed modified subtractive algorithm has been implemented in conjunction with an artificial intelligence techniques known as Adaptive Neuro Fuzzy Inference System (ANFIS) for judging the prediction accuracy of diabetic patients based on certain clinical tests. Finally, performance and accuracy of the technique has been presented and compared.**

Key Words-- Artificial Intelligence, ANFIS, MATLAB, Subtractive Clustering.

1. Introduction

DATA CLUSTERING is an approach for finding similarities in data and putting similar data into groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups [1]. The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever we are presented with a large amount of data, we usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Nevertheless, finding these groupings or trying to categorize the data is not a simple task for humans unless the data is of low dimensionality (two or three dimensions at maximum.) This is why some methods in soft computing have been proposed to solve this kind of problem. Those methods are called "Data Clustering Methods". Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. By finding similarities in data, one can represent similar data with fewer symbols for example. Also if one can find groups of data, one can build a model of the problem based on those groupings. Another reason for clustering is to discover relevance knowledge in data.

Usually the techniques presented in the present work are used in conjunction with fuzzy models. In particular, most of these techniques can be used as preprocessors for determining the initial locations for radial basis functions or fuzzy if then rules.

2. Literature Review

Junjie Caol, Kris Baoqian Panl, Kwok-Leung Tsuil, Shui-Yee Wongl (2012) [7]

Authors suggest a solution to the clustering problem is nearly perfect in one-dimensional space when the scan statistic is used. However, finding the solution is complicated by many problems in the higher-dimensional space. This is essentially a combinatorial optimization problem in the testing space. Exhaustive searching seems to provide a perfect solution, but this is computationally impossible for a large dataset. Some stochastic optimization methods have been developed to achieve an acceptable, global, optimized result. The most popular and widely used method (i.e. the spatial scan statistic) is based on the original scan statistic with reduced testing space by using a scan window with a fixed shape. Graph theory has been introduced into this area to aid in detecting arbitrarily shaped clusters. In this paper, we approach disease surveillance as a clustering problem, and we review some standard problem solving techniques. Furthermore, they propose a new statistic called the upper bound statistic to solve the clustering problem.

Mohammadreza Balouchestani, Sridhar Krishnan, (2014) [8]

Clustering and classification algorithms play an important role in long-term recording of ECG signals. The fundamental objective of their work was to establish new KMeans clustering algorithm based on CS theory. They have experimentally confirmed that our algorithm with a collaboration of PCA and K-NN as the dimensionality reduction and classification methods has demonstrated better performances than other methods. The proposed algorithm out-performed existing algorithms by increasing 11% classification accuracy in combination with PCA and K-NN methods in order to increase the classification speed. The proposed algorithm in combination with PCC and K-NN has achieved the following advantages: 1) 99.98% accuracy; 2) 99.98% ROC area; 3) 99.92% sensitivity. In addition, our algorithm in combination with PCA and PNN methods has demonstrated the following advantages: 1) high average accuracy for PNN classifier a 7 ; 2) 99.22% PPV; 3) 98.22% sensitivity. The benefit of using our algorithm can be divided into two areas. One area is to improve the wired ECG systems to wireless ECG. The second area of benefits emphasizes within increasing the efficiently of treatment for heat diseases.

Xiaohui Cui, Shi Liu and Likun Jia (2015) [9]
In their work the authors improved subtractive clustering algorithm. Firstly, they unified the fuzzy membership function based on AFS and SCM into a unified clustering framework which is called SDSCM. Then, the semantic concepts from user are used to automatically determine the density radius 1 and semi-automatically determine weight 2.The clustering number in SDSCM is not given in advance, which can be determined automatically by the given semantic concept from user and compression factor sqshFactor. The experimental results show that the clustering results of SDSCM are more semantic and the internal firmness of SDSCM is higher. Further, SDSCM has more practical significance and our future work is to apply the algorithm to the specific areas, such as improving retrieval performance, reforming browsing mechanism and finding similar documents.

Gautam Singh and Suresh Sundaram (2015) [10]
They have presented a maiden and successful methodology utilizing subtractive clustering for text-independent online writer identification. Over a variety of design parameters such as the radius f influence r a, squash factor 6' and number of prototypes M, we have demonstrated the behaviour of our subtractive clustering based approach. We demonstrated that this clustering scheme leads to comparable writer identification rates, with its traditional counterparts such as kmeans and fuzzy c-means clustering. This inference was drawn for two different scoring strategies thus bolstering our claim. Furthermore, we have observed that our proposed modified tf - idf scoring strategy outperforms the nearest prototype based scoring strategy. This work still leaves a lot of scope for future explorations. We mentioned that SC does not attempt to minimize the within-cluster error. This problem could be addressed in future by using SC clusters as seeds for k- eans or fuzzy c-means. Currently, in the SC algorithm, we have limited the number of prototypes to a fixed value M. The problem with variable number of clusters is that certain writers with higher number of prototypes can receive unduly higher score. This can happen for two reasons- (i) more terms in tf-sum in Equation 10 can inflate its value (ii) more cluster centres increase the chances of a sub-stroke in the test sample to get erroneously assigned to that writer. Although we solved this problem by fixing the number of prototypes, we could explore other ways as well. One way could be to assign weights to different prototypes of a given writer keeping number of prototypes variable as demanded by the conventional subtractive clustering. Furthermore, we could study the utilization of stroke-based features for testing the performance at line level. Dimensionality reduction techniques such as PCA and LDA on the derived 238-imensional features in this work may be attempted to possibly improve the identification rates. Notwithstanding these limitations and future research issues, this study provides a preliminary insight into the utility of subtractive clustering in writer identification domain.

3. Subtractive Clustering

Clustering is a process in which data are placed into clusters/groups, so that data in a cluster becomes similar to each other, and data in different clusters tend to be dissimilar. When the clustering estimation is applied to a group of input output dataset, each cluster centre can be considered as a fuzzy rule that describes the characteristic behaviour of the system. Each cluster centre corresponds to fuzzy rule, and the cluster identified represents the antecedent of this rule. This step forms the structure identification [3]. Clustering algorithms are used extensively not only to organize and do data categorization, but are also useful for compressing the data and model construction [4]. By finding data similarities, one can write similar data with less symbols. The density function for a data point is defined as the measure of potential for that data point. It is anticipated based on the distance of one data point from all other data points, Hence, a data point lying in a bundle of different data points will have a greater possibility of being a clustering centre, on the other hand a data point which is situated in an area of diffused and not concentrated dataset point will have a less possibility of being a cluster centre [5][6].
Subtractive clustering is a method to involuntarily generate fuzzy inference systems by finding clusters in input output training dataset. Subtractive clustering considers each data point as a potential cluster centre. The measure of potential for a data point is expected to be based on the distance of the data point from other data points. Hence, a data point located in a mound of different data points will have a greater chance of being a clustering centre, on the other hand a data point which is situated in an area of diffused and not concentrated data points will be having a least chance of being a cluster centre.
Next calculating the potentials of each data point, the data point with the greatest potential value is elected as the first clustering centre. To locate the next clustering centre, data points potentials must be repeated. For every data point, an amount relative to its distance to the first cluster centre will be deducted. This minimises the possibility of a data point centre. After revising the potential of every data points, the data point with the greatest potentials will be chosen after that cluster centre. The capability of data points in the initial step is measured as given by Chiu [2].
Given a collection of n data points $\{x_1, \ldots\ldots x_n\}$, the subtractive clustering algorithm considers each data point as a potential cluster center. A density measure at a data point x_i is defined as

$$D_i = \sum_{j=1}^{n} e^{-\left|x_i - x_j\right|^2 / (r_a/2)^2} \qquad (1)$$

where the cluster radius r_a is a positive constant. Thus, a data point that has many neighbouring data points will have a high potential of being a cluster center. The radius r_a defines a neighbourhood. Data points beyond this radius have little effect on the density measure. The choice of r_a plays an important role in determining the number of clusters. Large

values of r_a will generate a limited number of clusters, while small values of r_a will generate a large number of clusters. After the density measure of each data point has been calculated, the first cluster center is chosen to be the data point with the maximum density value. Suppose x_{c_1} is the point selected and D_{c_1} is its density measure, then the density measure for each data point x_i is revised by the formula

$$D_i = D_i - D_{c_i} e^{-\left|x_i - x_{c_i}\right|^2 / (r_b/2)^2} \quad (2)$$

where r_b is a positive constant. The data point near to the very first cluster center x_{c_1} will have significantly reduced density measures, so that they are very less likely to be selected as the following clustering center. The constant r_b is as a rule larger than r_a to prevent closely spaced cluster centers. Generally r_b is specified as 1.5 times of r_a. After revising the density measure for each data point, the next cluster center x_{c_2} is selected, and all of the density measures for data points are revised again. Subtractive clustering can be used as a standalone approximate clustering algorithm in order to estimate the number of clusters and their locations.

3.1 Limitations of Original Algorithm

Some of the limitations of the main algorithm are, firstly, one should know the number of cluster right at the beginning because this algorithm does not notify about accuracy of the number of clusters, hence one this may lead to wrong selection of the number of cuaters L. For e.g. data might actually have ten clusters but one ha taken L= 8, then algorithm will give only 8 clusters. These clusters will be far apart and not compact enough.. On the other hand if one choses the number of clusters chosen less then it may lead to putting dissimilar points in one cluster. Secondly, it may lead to dead unit problem, i.e. if cluster center is incorrectly selected, it may not be updated at all and thus will not represent a class. Finally, it converges to restricted minima, i.e. for chosen points

3.2 Proposed Modified Algorithm

Input: Dataset D of N data points (i = 1 to N) ; Number of clusters wanted = L

Output: N data points clustered into L cluster.

Phase 1 Steps :

1. Adjust m = 1;
2. calculate the distance between each data points and all remaining data points in the set D;
3. Search for the nearest pair of data points from the set D and develop a data point set A_m (1<= m<= L) which has these two data points, eliminate these two data point from the set D;
4. Locate the data point in D that is nearest to the data point set A_m, Add it to A_m and delete it from D;
5. replicate step 4 till the number of data points in A_m reaches 0.75*(N/L);
6. If m<L, then m = m+1, search for further pair of data points in D such that the distance measure is the least, develop another data point set A_m and eliminate them from D, next to step 4;
7. For every data points set A_m (1<=m<=L) find the arithmetic mean of the vectors of data points in A_m. These will be the initial centroid.

Phase 2 Steps :

1. calculate the distance between every data point d_i (i less than equal to N and greater than equal to 1) to all the centroids C_j (j less than equal to L and greater than equal to 1) as $d(d_i, C_j)$;
2. For every data point d_i, locate the nearby centroid C_j and allocate d_i to cluster *j*.
3. Adjust Cluster Id[i]=j; // j:Id of the neighboring cluster
4. Set Closest _Dist[i]= $d(d_i, C_j)$;
5. For every cluster *j* (1<=j<=L), remeasure the centroids;
6. Repeat
7. For every data-point d_i,
   a. calculate its distance from the centroid of the current adjoining cluster;
   b. If this distance is <= to the present nearest distance, the data point stays in the cluster;
   c. otherwise for each centroid *cj* (j less than equal to L and greater than equal to i) calculate the distance $d(d_i, C_j)$;
   d. End for;
8. allocate the data point d_i to the cluster with the adjacent centroid C_j
9. Set ClusterId[i]=j;
10. Set Nearest_Dist[i] = $d(d_i, C_j)$;
11. End of (step(2));
12. For every cluster *j* (j less than equal to L and greater than equal to i), remeasure the centroids till the convergence criteria is reached.

4. Results and Discussions

These algorithms have been tested on diabetes real datasets obtained from UCI machine repository. It has 8 attributes, 768 instances, and 2 classes, hence L=2 has been taken as the number of clusters. The same has been tested on MATLAB R2009b software. The above datasets have been pre-processed by applying correlation analysis in order to find out the maximum correlated attributes and duplicate ones are deleted from dataset and thus clustering is carried out on processed data. Performance Analysis is carried out based on the number of iterations and Root Mean Square Error. The input and output data values used for generating the ANFIS prediction model are graphically represented in Figure 1 below. Here for generating the model 499 datasets has been used for training the proposed model and the remaining 269 datasets have been

International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

used for checking the model accuracy. The training and testing of the model has been carried out for 200 epochs.

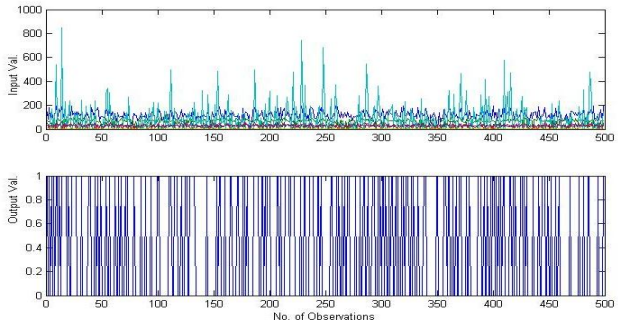


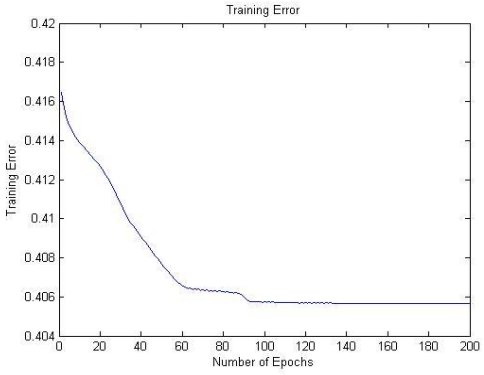
Fig. 1. Plot of Input and Output Data values

The model parameter values used for during testing the prediction accuracy of diabetic patients using both original and modified algorithms is given below in Table:1.

Table 1: Model Parameter Values Used

	For Original Algo	Modified Algo
Number of nodes	90	22
Number of linear parameters	40	8
Number of nonlinear parameters	70	12
Total number of parameters	110	20
Number of training data pairs	499	499
Number of checking data pairs	269	269
Number of fuzzy rules	5	2

From Table 1 it is seen that modified algorithm led to the reduction in the number of fuzzy rules generated during the model development process.



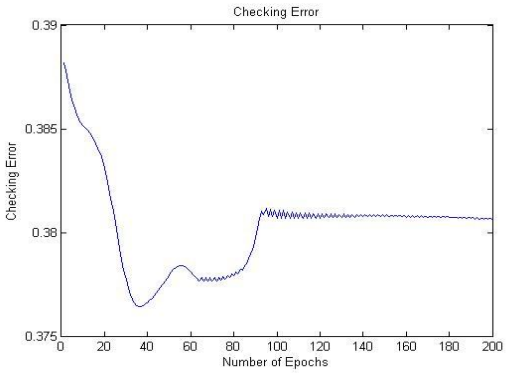


Fig.2 : Plot of RMSE Value during training and testing phase using Modified Algorithm

Figure 2 shows the plot of RMSE values during training and testing phase of model development using modified algorithm. From the graph it is seen that during training phase upto 90 epochs there was a reduction in the RMSE value after which the value saturated to 0.4165, while on the other hand during testing phase initially there was a sharp fall in RMSE value which again shoed an increasing trend, which later on saturated to 0.3831. Table 3 given below shows the RMSE values using both the original and the modified algorithm in case of training and testing datasets and their plots are given in Figure 3 below.

Table 2: Comparative Results of Original and Modified Subtractive Clustering Algo

		RMSE			
		Original Algo		Modified Algo	
		Trg. Data	Chk. Data	Trg. Data	Chk. Data
1	System generated using ANFIS	0.3862	0.4060	0.4165	0.3882
2	Optimised Output value	0.3830	0.4156	0.4127	0.3831
3	Testing the overfitting	0.3858	0.4011	0.4096	0.3764

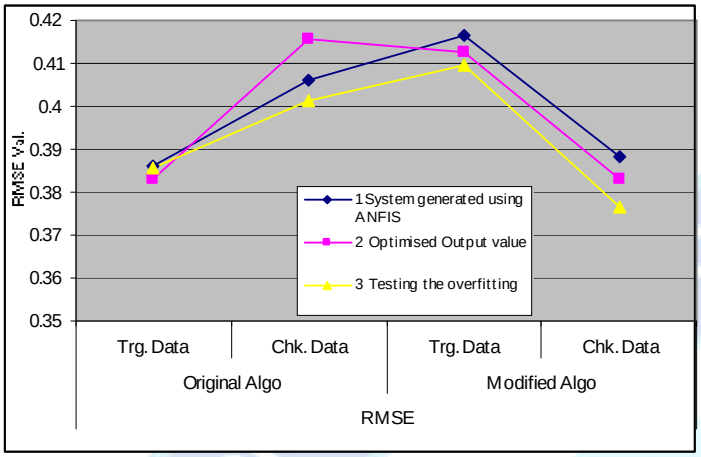


Fig. 3. Comparative graphical plot of RMSE values using Original and Modified Algorithm

From the perusal of data given in Table3 it is clear that there is a reduction in the RMSE values for checking datasets using modified algorithm as compared to original algorithm, although the model training values are a bit higher, which again shows that there is less of over-fitting during the training process of the model generation.

5. Conclusion

From the perusal of the data given in the above tables and analytical study of the figures it is evident that although the training RMSE values for training datasets for the original algo is slightly better than the modified algo but during the testing phase of the model the RMSE results have improved with the modified algo. This clearly demonstrates that the modified subtractive clustering algo has resulted in better model development for prediction.

References:

[1] Jang, J.-S. R., Sun, C.-T., Mizutani, E., "Neuro-Fuzzy and Soft Computing – A Computational Approach to Learning and Machine Intelligence," *Prentice Hall.*

[2] Azuaje, F., Dubitzky, W., Black, N., Adamson, K.,"Discovering Relevance Knowledge in Data: A Growing Cell Structures Approach," *IEEE Transactions on Systems, Man, and Cybernetics- Part B: Cybernetics, Vol. 30, No. 3, June 2000 (pp.*448)

[3] VAIDEHI, V., MONICA, S., MOHAMMAD SHEIKH SAFEER, S., DEEPIKA, M. AND SANGEETHA, S., (2008), "A Prediction System Based on Fuzzy Logic", *Proceedings of World Congress on Engineering and Computer Science.*

[4] HAMMOUDA, K. A., "Comparative Study of Data Clustering Techniques".

[5] CHIU, S., (1994), "Fuzzy Model Identification based on cluster estimation", *Journal of Intelligent and Fuzzy Systems*, 2 (3), pp 267–278.

[6] PRIYONO, A. RIDWAN, M., et. al. (2005), Generation of fuzzy rules with subtractive clustering", *Journal Teknologi.*, 43(D), . pp 143-153.

[7] Junjie Caol, Kris Baoqian Panl, Kwok-Leung Tsuil, Shui-Yee Wongl,(2012) "Disease Surveillance by Clustering Based on Minimal Internal Distance"978-1-4577-1911-0/12 IEEE

[8] Mohammadreza Balouchestani, Sridhar Krishnan, (2014) " Fast Clustering Algorithm for Large ECG Data Sets Based on CS theory in Combination with PCA and K-NN Methods" 978-1-4244-7929-0/14©2014 IEEE

[9] Xiaohui Cui, Shi Liu and Likun Jia (2015)"An Improved Method of Semantic Driven Subtractive Clustering Algorithm" 978-1-4799©2015 IEEE

[10] Gautam Singh and Suresh Sundaram (2015) "A Subtractive Clustering Scheme for Text-Independent Online Writer Identification" 978-1-4799-1805-8©2015 IEEE.