4

Data Quality Enhancement Technology to Improve Decision Support

Ahmad Shahi, Rodziah binti Atan and Nasir bin Sulaiman
University Putra Malaysia
Malaysia

1. Introduction

Uncertainty is a very important aspect of real human life. It means "Not knowing with certainty; such as cannot be definitely forecast" [1]. This uncertainty in fuzzy systems occurs mainly due to volume of work, lack of theoretical knowledge and lack of experimental results [2].
Mendel [3] has noted that uncertainty also exists while building and using typical fuzzy logic systems. He has described four sources of uncertainty: *Uncertainty about the meanings of the words that are used in a rule*, this is the uncertainty with the membership functions because membership functions represent words in a fuzzy logic system. It can be both antecedents and consequents; *Uncertainty about the consequent that is used in a rule,* this is the uncertainty with the rule itself. A rule in FLS describes the impact of the antecedents on the consequent. Expert may vary in their opinion to decide this nature of impact; *Uncertainty about the measurements that activate the FLS,* this is the uncertainty with the crisp input values or measurements that activates the FLS. These measurements may be noisy or corrupted. The noise can again be in a certain range or totally uncertain meaning stationary or non-stationary; *Uncertainty about the data that are used to tune the parameters of a FLS*, this is the uncertainty with the measurements again.
To deal with the uncertainty, fuzzy logic is a proper way to model human thinking. Although it was introduced by Lotfi Zadeh in 1965, it has been used to build expert systems for handling ambiguities and vagueness associated with the real world problems which involve different kinds of uncertainty [4].Thus in order to strengthen fuzzy system model, quality of data as an input of the model should be enhanced. Outliers and noisy data, these uncertainty arise from mechanical faults, change in system behavior, fraudulent behavior, network intrusions, sensor and device error, human error and so on [5, 6]. However, to strengthen fuzzy system model, outliers should be isolated that, the following section demonstrates about details of isolating outliers.

1.1 The reason of isolating outliers

The main reason for isolating outliers is associated with data quality assurance. The exceptional values are more likely to be incorrect. According to the definition, given by Wand and Wang [7], unreliable data represents an unconformity between the state of the database and the state of the real world. For a variety of database applications, the amount of erroneous data may reach ten percent and even more [8]. Thus, removing or replacing

outliers can improve the quality of stored data. Isolating outliers may also have a positive impact on the results of data analysis and data mining. Simple statistical estimates, like sample mean and standard deviation can be significantly biased by individual outliers that are far away from the middle of the distribution. In regression models, the outliers can affect the estimated correlation coefficient [9]. Presence of outliers in training and testing data can bring about several difficulties for methods of decision-tree learning, described by Mitchell in [10] and parameters in Gaussian membership function parameters in [2]. For example, using an outlying value of a predicting nominal attribute can unnecessarily increase the number of decision tree branches associated with that attribute. In turn, this will lead to inaccurate calculation of attribute selection criterion (e.g., information gain). Consequently, the predicting accuracy of the resulting decision tree may be decreased. As emphasized in [11], isolating outliers is an important step in preparing a data set for any kind of data analysis.

1.2 Effective quality of data on technology of fuzzy system

Fuzzy systems are expressed by membership functions. The outlier and noise are kinds of uncertainty which have effect on the membership function parameters, such as the Gaussian membership. In Gaussian, there are two parameters, mean and standard deviation, which are tuned based on the dataset. However, if the desired data is extracted from the dataset, Mean and Standard deviation can be accurate parameters for the Gaussian membership. Hence, to make a robust model, the outliers must be detected and the noisy data must be removed from the dataset.

There is a direct, although rarely explored, relation between uncertainty of input data and fuzziness expressed by Membership Functions (MFs). Various assumptions about the type of input uncertainty distributions change the discontinuous mappings provided by crisp logic systems into more smooth mappings that are implemented in a natural way by fuzzy rules using specific types of MFs. On the other hand shifting uncertainty from fuzzy rules to the input values may simplify logical rules, making the whole system easier to understand, and allowing for easy control of the degree of fuzziness in the system [12].

If regions of the data of different classes are highly overlapping or if the data is noisy, the values of the membership degrees could be misleading with respect to rule confidence if the core region is modeled too small. In fact, we show that data regions with a high membership degree need not to be the regions with a high rule confidence. This effect that we call membership is unrobustness [2].

Therefore, the Fuzzy C-Mean clustering (FCM) is utilized to detect the outlier and statistic equation is used to remove the noisy data in order to improve the quality of the data.

2. Design of method

As shown in Figure 1, the Weka software which was developed at Waikato University [13], is used for pre-processing the data in the dataset. After cleaning the data, the FCM with statistic equation (which is described in following section) were utilized to detect outliers, remove noisy data and extract the desired data to get data of high quality. In the next step, Type-1 FLS with gradient descent algorithm) were used to make a decision on such data, after analyzing the data to decide on the parameters to be used, including temperature, humidity and so on. The important part of this technique is that the gradient descent

algorithm was used to tune the membership function parameters. The proposed method can be seen in Figure 1. The details of the proposed method are described in following sections.

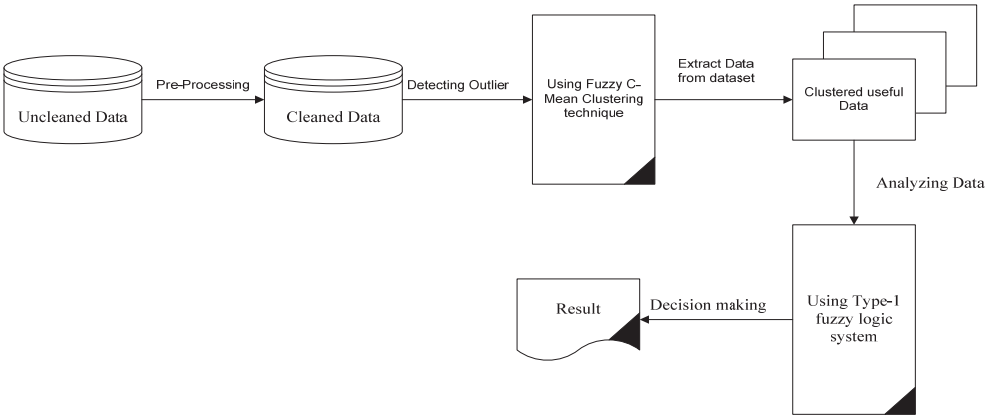


Fig. 1. Flowchart of Proposed method

2.1 Fuzzy C-Mean clustering

In fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to one cluster. To manage uncertainty in getting data quality and accurate model, a new method known as the Fuzzy C-Mean (FCM) clustering is needed.

The FCM computes the distance measured between two vectors, where each component is a trajectory (function) instead of a real number. Thus, this Fuzzy C-mean clustering is rather flexible, moveable, creatable, as well as able to eliminate classes and any of their combination. From the huge number of clustering methods, the fuzzy clustering was focused on in the methodology of the present study since the degree of membership function on an object to the classes found provides a strong tool for the identification of changing class structures. Thus, the Fuzzy C-Means is used in order to build an initial classifier and to update the classifier in each cycle; nevertheless, the methodology presented can still be extended to any other techniques which determine such degrees of membership (e.g. probabilistic clustering, etc.) [14].

Before the FCM could be utilized, the noise was removed from the dataset due to affect on clustering data. Based on the statistic definition:

According to [15] a noise is considered to be more than three standard deviations away from the mean which is formulated as below:

$$\text{Noises} = \text{abs}(\text{object} - \text{MeanMat}) > 3*\text{SigmaMat};$$

In which, Meanmat is mean, SigmaMat is Standard deviation and abs is the functioning in mathematic for Absolute value, i.e. instance: abs (-5) =5, which was implemented in MATLAB software.

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. The method developed by [16, 17] is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , 1 \le m < \infty \tag{1}$$

Where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension centre of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the centre.
Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centre c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}}, \tag{2}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{3}$$

This iteration will stop when the $\max_{ij}\left\{\left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, and k is the iteration step. This procedure converges to a local minimum or a saddle point of J_m.
The algorithm is composed of the following steps [17]:

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
2. *At k-step: calculate the centre vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}}$$

4. *If* $|| U^{(k+1)} - U^{(k)} || < \varepsilon$ *then STOP; otherwise return to step 2.*

The Fuzzy C-Mean clustering and statistic equation were implemented in the MATLAB software.

2.2 Type-1 fuzzy logic system

Fuzzy logic was developed by Lotfi Zadeh a professor at the University of California, Berkley. Fuzzy system is useful for real world problems where there are different kinds of uncertainty [18]. The idea of fuzzy logic was to show that there is a world behind conventional logic. This kind of logic is the proper way to model human thinking. Fuzzy logic is recently getting the attention of artificial intelligence researchers. It is being used to build expert systems for handling ambiguities and vagueness associated with real world problems. Figure 2 shows the architecture of Type-1 Fuzzy system with gradient descent algorithm.

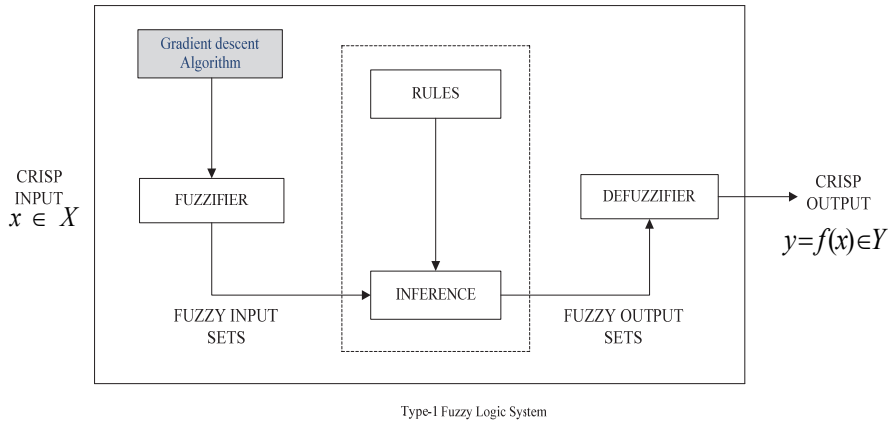


Fig. 2. Structure of Type-1 Fuzzy Logic system with Gradient Descent Algorithm

Gradient descent Algorithm: Gradient descent technique is a training algorithm which was used to tune the membership function parameters in fuzzy system. Using this algorithm the membership function parameters can be optimized and the error rate is reduced to get more accurate results. The details of the algorithm were explained in [19].

Fuzzification Process: According to [20] fuzzifying has two meanings. The first is the process fining the fuzzy value of a crisp one. The second is finding the grade of membership of a linguistic value of a linguistic variable corresponding to a fuzzy or scalar input. The most used meaning is the second. Fuzzification is done by membership functions.

Inference Process: The next step is the inference process which involves deriving conclusions from existing data [20]. The inference process defines a mapping from input fuzzy sets into output fuzzy sets. It determines the degree to which the antecedent is satisfied for each rule. These results in one fuzzy set assigned to each output variable for each rule. MIN is an inference method. According to [21] MIN assigns the minimum of antecedent terms to the matching degree of the rule. Then fuzzy sets that represent the output of each rule are combined to form a single fuzzy set. The composition is done by applying MAX which corresponds to applying fuzzy logic OR, or SUM composition methods [20].

Defuzzification Process: Defuzzification is the process of converting fuzzy output sets to crisp values [20]. According to [22] there are three defuzzification methods used : *Centroid*, *Average Maximum* and *Weighted Average*. Centroid method of Defuzzification is the most commonly used method. Using this method the defuzzified value is defined by:

$$\text{Centroid}=\frac{\int x\mu(x)dx}{\int \mu(x)dx} \tag{4}$$

Where $\mu(x)$ is the aggregated output member function. The details of Fuzzy system have been explained in [20] [23].

3. Experiments and results

The dataset for preprocessing is taken from Politecnico in Torino (Italy) that contains two attributes, temperature and humidity. These measures have been taken from a sensor and recorded by a computer at regular intervals of about 15 minutes (the average interval was estimated at 896 seconds) [24]. The dataset has been preprocessed by WEKA software to check missing value. And the statistic equation was utilized to remove noisy data, after that FCM clustering was used to detect outliers. By detecting outliers, desired data was extracted from dataset as an input of fuzzy system for controlling and decision making.
The Italy dataset possesses 4600 instances. Based on the equation coded in the Matlab, the noise was found and removed from the dataset. The graphic view of the Italy dataset before removing noise is depicted in Figure below.

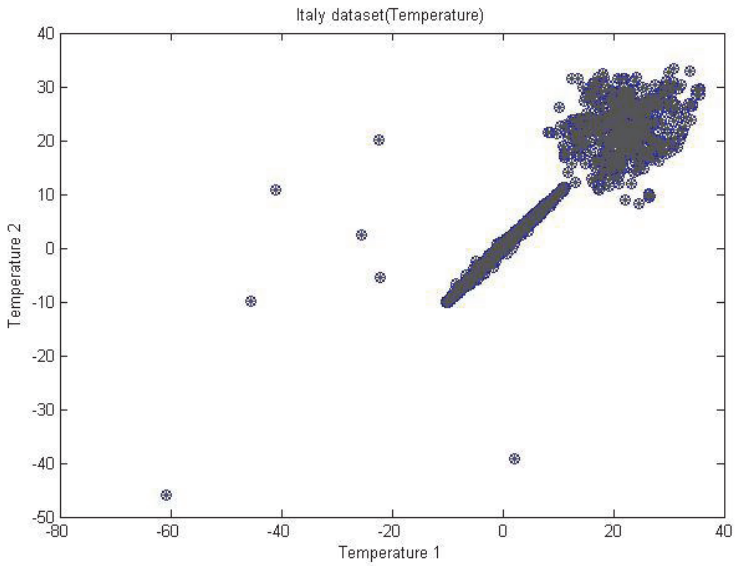


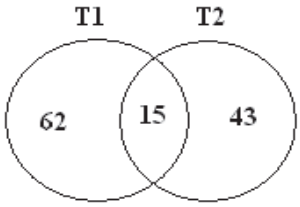
Fig. 3. Graphical View of the Original Italy Dataset

As shown in Figure 3, the attributes (temperature 1 and 2) were derived from the device sensor. The extracted data may include noisy data due to device and measurement errors. This would decrease the quality of the data. Therefore, the program in this study was applied based on the definition of the statistic equation on the data to remove the noisy data. The characteristics and results are shown in Table 1.

Attribute	Number of instances with noise	Number of noisy instances	Number of common noisy instances	Total noise	Number of instances without noise
Temperature 1	4600	62	15	90	4510
Temperature 2	4600	43			4510

Table1. The Number of the Italy dataset instances with and without noise

Table 1 shows that the number of the noisy data in the attribute Temperature 1 is 62 and this is 43 for the attribute Temperature 2; 15 noise instances are common among the attributes Temperature 1 and Temperature 2, and thus the total instances for two attributes after removing noise is 4510. It means:



4600-(62+43)-15=4510

After removing the noise from the dataset, the Fuzzy C-Mean clustering (FCM) was used to detect the outliers and extract the desired data.

The FCM was used to cluster the data after removing the noisy data to mine the desired cluster. The noise detected in the present study was found to have effect on the FCM, and the mean is not defined well due to the noise. Therefore, the noisy data must be removed in order to get the accurate results.

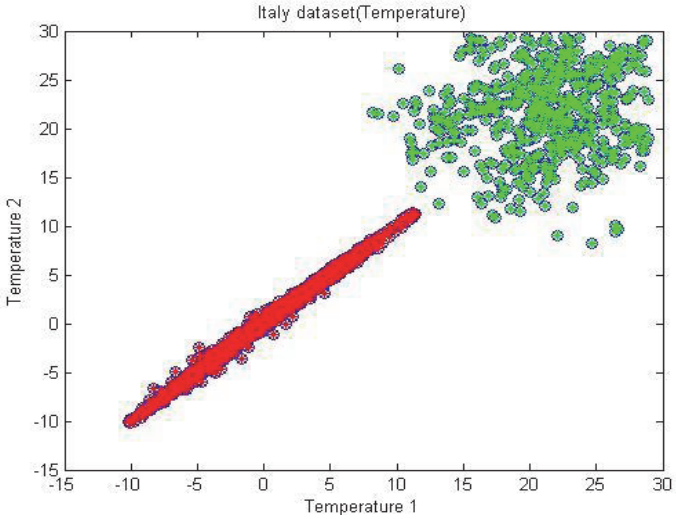


Fig. 4. Detecting and Clustering Outliers on Italy Dataset.

The attributes, which were derived after removing noise and clustered based on the FCM algorithm, are shown in Figure 4. The number of clusters is 2 (k=2, k is a number of cluster); green represents one cluster and red is another cluster which possesses behaviour that is different from each other.

Dataset	Number of cluster	Number of Instances with noise	Number of noisy instance	Number of instance without noise	Number of instances		Center of Cluster (Mean)	
					Cluster 1	Cluster 2	Cluster 1 (x,y)	Cluster 2 (x,y)
Italy	2	4600	90	4510	566	3944	(20.9173, 21.4507)	(0.7194, 0.7179)

Table 2. Summary of characteristics the Italy dataset

As shown in Table 2, are characteristics of Italy dataset. The dataset has 2 clusters; cluster 1 contains 566 records and cluster 2 has 3944 records. Based on the two dimensions, the centre of cluster 1: Temperature1=20.9173 and Temperature2=21.4507; likewise for cluster 2: Temperature1=0.7194 and Temperature2=0.7179.

3.1 Analysis using type-1 FLS

The dataset was pre-processed using the Weka to check for any missing values. After that, the data was used as an input for the method (i.e. Type-1 Fuzzy Logic System).
In Figure 5, blue lines show the release desired or the real data and red lines are obtained from the data predicted as the outcome of the method (Type-1 FLS).

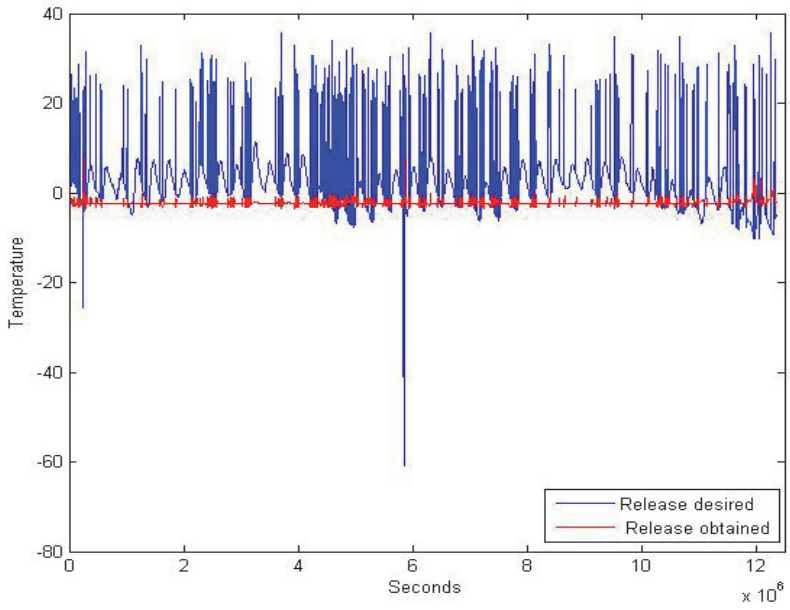


Fig. 5. Type-1 FLS on Italy dataset

Based on Figure 5, the release desired colored in blue (real data) and release obtained colored in red (predicted data) are not close to each other. The difference is due to the low quality of the data that contains noise and outliers, affecting the membership function and causing the function to be inaccurate and not robust. The membership function has two parameters, namely the mean and standard deviation, which are tuned based on the data.

3.2 Analysis and results using proposed method

After removing the missing value, FCM with statistic equation are applied to detect outliers and remove noisy data on the Italy dataset, the desired clustered data was extracted and entered as an input entry into the Type-1 Fuzzy Logic System with gradient descent algorithm to tune membership function parameters. The result shows a different type of graph that presenting the method effect to the dataset.

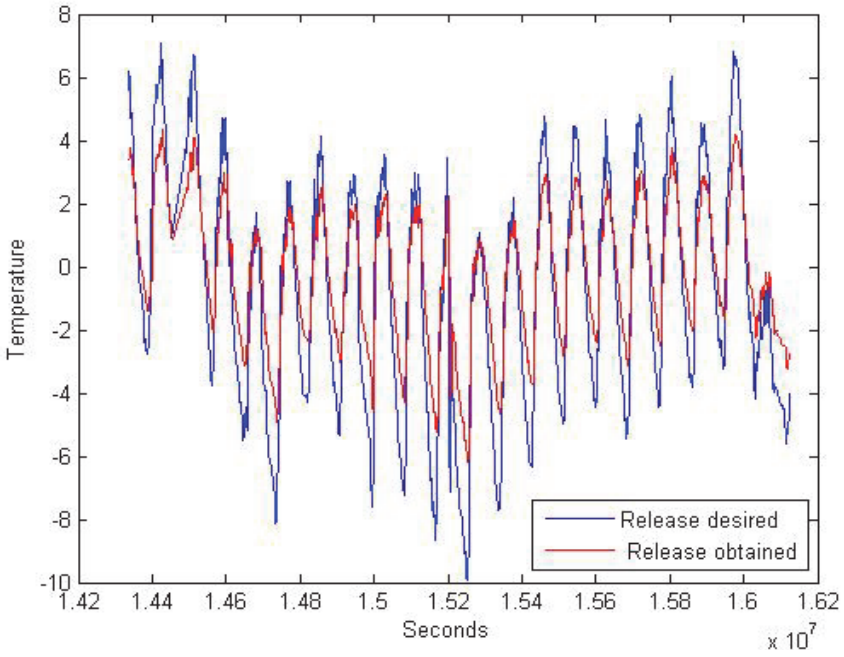


Fig. 6. Proposed Method on Italy Dataset

Figure 6 shows the output of Proposed. The attribute is a temperature based on time (second). A small scale of data (desired clustered data) was used to show the system behaviour (Proposed Method).

Blue shows the desired data or the real data and red is the obtained or predicted data. If data before the contribution (Type-1 FLS) were compared, it resulted much better and represents predicted data more closely to the real ones.

3.3 Accuracy measurements

To evaluate the results we used standard measurement called Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) which formulated as below:

If O is the actual observation for time period t and P is the forecast for the same period, then the error is defined as:

Error: $$e_t = O_t - P_t \tag{5}$$

Since there are observations and predictions for n time periods, then there will be n error terms, and the following standard statistical measures can be defined:

Mean Absolute Error: $$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t| \tag{6}$$

Root Mean Square Error (RMSE): $$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}e_t^2} \tag{7}$$

The MAE is defined by first making each error positive by taking its absolute value and then averaging the results. The RMSE is a quadratic scoring rule which measures the average magnitude of the error. The equation for the RMSE is given in both of the references. Expressing the formula in words, the difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable.

The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude. Both the MAE and RMSE can range from 0 to ∞. They are negatively-oriented scores: Lower values are better. Each of these statistics deals with measures of accuracy whose size depends on the scale of the data [25].

Dataset: Italy (Temperature)	Before Contribution	After Contribution
Measurement metrics/methods	Type-1 FLS	Proposed Method
MAE	7.0321	1.3994
RMSE	10.5019	1.6590

Table 3. Accuracy Measurements of Italy Dataset

In such circumstances, Table 3 shows results of error value for type-1 FLS before and after contribution .For MAE measurement is 7.0321 and 1.3994 using type-1 FLS before and after clustering respectively. These values in RMSE measurement are 10.5019 and 1.6590. The result shows the method after contribution achieve minimum error rate in both MAE and RMSE. RMSE

These effective results are due to the mechanism of Type-1 FLS* with FCM* method. Type-1 FLS$^+$ improves the quality of data by detecting outliers, removing noisy data and tuning MFs parameters by training algorithm is called Gradient Descent algorithm. Thus this method is a significant technique that can improve the accuracy of weather situation.

4. Conclusions

Outlier and noise are part of uncertainty that arises due to mechanical faults, changes in system behavior, fraudulent behavior, network intrusions, human errors, keyboard error, hand writing error and so on that affect on measurement of Gaussian membership function parameters. In Gaussian there are two parameters, Mean and Standard deviation that are tuned based on dataset, therefore if we do not extract useful knowledge or desired clustered data from dataset, Mean and Standard deviation will not be accurate parameters for Gaussian membership function. From the huge number of clustering methods, Fuzzy C-Mean clustering is flexible, moveable, creatable, elimination of classes and any their combination. Since the degree of membership function on an object to the classes found provides a strong tool for the identification of changing class structures. Fuzzy C-Mean in order to build an initial classifier and to update our classifier in each cycle, thus we utilized Fuzzy c-mean clustering with statistic equation to remove noisy data and detect outlier and mine valuable data to get accurate result with Type-1 Fuzzy Logic Systems and gradient descent algorithm.

By applying proposed method, the quality of data has been improved (As shown in Table 3). The proposed method enhanced the data quality. Thus, by improving the quality of data, the accurate decision making will be achieved in decision support system.

5. References

[1] Russell, G., *Grossett:"Webster's New Dictionary and Thesaurus"*. Geddes and Grosset Ltd., New Lanark, Scotland, 1990.

[2] Rahman, A., *Handling imprecision and uncertainty in software quality models.* 2005.

[3] Mendel, J.M., *Uncertain rule-based fuzzy logic systems: introduction and new directions*. 2000: Prentice Hall.

[4] Zadeh, L.A., *The concept of a linguistic variable and its application to approximate reasoning*. Information sciences, 1975. 8(3): p. 199-249.

[5] Last, M. and A. Kandel. *Automated detection of outliers in real-world data*. 2001: Citeseer.

[6] Cherednichenko, S., *Outlier detection in clustering.* 2005.

[7] Wand, Y. and R.Y. Wang, *Anchoring data quality dimensions in ontological foundations.* 1996.

[8] Wang, R.Y., M.P. Reddy, and H.B. Kon, *Toward quality data: An attribute-based approach.* Decision Support Systems, 1995. 13(3-4): p. 349-372.

[9] Clarke, E. and T. Coladarci, *i> Elements of Statistical Reasoning.</i> New York*. 1999, Wiley.

[10] Mitchel, T., *Machine learning.* Machine Learning, 1997. 48(1).

[11] Pyle, D., *Data preparation for data mining*. 1999: Morgan Kaufmann Pub.

[12] Rehm, F., F. Klawonn, and R. Kruse, *A novel approach to noise clustering for outlier detection.* Soft Computing-A Fusion of Foundations, Methodologies and Applications, 2007. 11(5): p. 489-494.

[13] Kirkby, R. and E. Frank, *WEKA Explorer User Guide for Version 3-5-3.* University of Waikato,(June 2006), 2006.

[14] Crespo, F. and R. Weber, *A methodology for dynamic data mining based on fuzzy clustering.* Fuzzy Sets and Systems, 2005. 150(2): p. 267-284.

[15] Macfie, B.P. and P.M. Nufrio, *Applied statistics for public policy.* 2006: ME Sharpe Inc.

[16] Dunn, J.C., *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.* Cybernetics and Systems, 1973. 3(3): p. 32-57.

[17] Bezdek, J.C., *Pattern Recognition With Fuzzy Objective Function Algoritms, Plenum Press.* New York, 1981.
[18] Zadeh, L.A., *Outline of a new approach to the analysis of complex systems and decision processes.* IEEE Transactions on Systems, Man, and Cybernetics, 1973. 3: p. 28-44.
[19] Li-Xin, W., *A course in fuzzy systems and control.* prentice Hall, 1997.
[20] Siler, W. and J.J. Buckley, *Fuzzy expert systems and fuzzy reasoning*. 2005: Wiley-Interscience.
[21] Hwang, G.J. and S.T. Huang, *New environment for developing fuzzy expert systems.* J INF SCI ENG, 1999. 15(1): p. 53-69.
[22] *http://www.jimbrule.com/fuzzytutorial.html*. Last accessed 4 November 2006.
[23] *http://www.urc.bl.ac.yu/manuals/adv/fuzzy.* Lased accessed 3 November 2006.
[24] Mencattini, A., et al. *Local meteorological forecasting by type-2 fuzzy systems time series prediction*. 2005.
[25] Mountis, A. and G. Levermore. *Weather prediction for feedforward control working on the summer data*. 2005.