Int. Journal of Math. Analysis, Vol. 6, 2012, no. 59, 2911 - 2920

Noise Filtering in Unsupervised Clustering Using Computation Intelligence

Chaloemchai Lowongtrakool

Department of Computer Science, Faculty of Science
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
s9062914@live.kmitl.ac.th

Nualsawat Hiransakolwong

Department of Computer Science, Faculty of Science
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand

Abstract

The present data used for clustering contains data record not related to processing. This may be spam or noise causing error and delay of processing. However, it is necessary to process this kind of data together but the results may be incorrect, depending on the quantity of noise. Therefore, it will be much better if data cleaning is conducted before processing system data. This paper proposes a method to develop unsupervised clustering intelligence to reduce the quantity of spam. This computational intelligence system applies the first layer of radial basis function network as an input layer of the system for incremental work. The results of system can provide level of accuracy of least related membership in any cluster which is called the data not relevant to the dataset or the data not associated with the dataset. According to an experiment, the data from UCI machine learning repository was used to test the system efficiency while classification algorithm in a program called weka3.6 was also utilized to test the system accuracy. The results from noise filtering help the data processing more precise, compared to the processing without the noise filtering.

Keywords: Noise Filtering, Unsupervised Clustering, Computation Intelligence

1 Introduction

The noise filtering or data filtering is conducted to eliminate the irrelevant content from the real data. This procedure is very significant but it is nearly impossible to get rid of the noise from all real data. Moreover, it is the first thing to consider when using the data for processing requiring accuracy and precision. The noise may be unfavorable thing, especially for the data used for unsupervised clustering. In this case, the data obtained from an input is raw data without any classifying. The system needs to learn from the raw data itself by applying various current techniques which is absolutely challenging. However, a problem is the system will learn and make its own decision from received data including both correct and incorrect data caused from the noise. If the noise is not eradicated, the system cannot realize the noise or irrelevant data which affects wrong decision making. In previous researches [1 8 14 17], a neural network was employed to remove the noise and train the dataset containing correct class. Then the values of value was tested and compared. If they were unmatched or there were some errors, such data would be regarded as the noise. Furthermore, truth value and false membership were another ways to compare the real result and the outcome of the system testing. If the difference was found, such record would be considered as the noise**.**

2 Related works on noise filtering

2.1 Complementary Neural Network (CMTNN)

CMTNN [13] is a technique based on ANNs. It concentrates on class noise or misclassification error [1 19]. Besides, the truth and false membership values are compared [15 16] during the training. If the difference is found, they will be regarded as the noise. The procedure is described as below.

Step 1) The values of Truth Neural Network and Falsity Neural Network are trained by using both membership truth value and membership false under value feedforward backpropagation neural network.

Step 2) A prediction of value (Y) during the data training by using a target value (T) provides the outcome of both values (O) comparison as Truth NN and Falsity NN. These values will be regarded as the noise when the results of predictive values and calculated values are different. For example, if the outcome of target during the training of Truth NN is 0, the complement value of target during the training of Falsity NN must be 1.

In case of Truth NN, if $Y_{\text{Truth i}} \neq O_{\text{Truth I}}$, then

$$M_{\text{Truth}} \leftarrow M_{\text{Truth}} \; U \; \{T_i\} \tag{1}$$

In case of Falsity NN, if $Y_{\text{Falsity i}} \neq O_{\text{Falsity i}}$, then

$$M_{\text{Falsity}} \leftarrow M_{\text{Falsity}} \, U \, \{T_i\} \quad (2)$$

Step 3) A set received from train(T_c) will lead to data cleaning stemmed from the comparison between the values of Truth NN and Falsity NN as shown in this below formula.

$$T_c \leftarrow T\text{-}(M_{\text{Truth}} \cap M_{\text{Falsity}}) \quad (3)$$

According to an intersection between M_{Truth} and M_{Falsity}, the noise is the different values.

2.2 Automatic noise reduction (ANR)

ANR [18] is based on multi-layer neural network trained by Standard Backpropagation. In addition, a class of answer is applied to adjust the value of learning to be similar to the network. Later, the answer class will be trained. If there is any change in a label class, it will be considered as the noise or mislabeled [2 3 7] and the noise record will be removed from the dataset. In the stage of training of Backpropagation Network, the values of weight and error rate is adapted to meet the least error value. It can be said that the mislabeled value is at the lowest level, too. There are 3 reasons of this technique to remove the noise from the dataset as follows.

Reason 1) The error rate is measured in the stage of training. If the error rate is found at low level (e.g.<30%), it will be considered as the noise.

Reason 2) In case of wrong mislabeled pattern, if the accuracy is measured and found that there are 2 possible class values, they will be regarded as the noise.

Reason 3) The data cleaning is required if any data affects classifier quality to train the system, except small dataset.

Regarding to the former theories, it is essential to use the answer class for the training to figure out the answer value. Moreover, the truth and false values can be applied to the comparison. To the training of neural network, if the correct data is used for the training, the results will be precise as well. Nevertheless, this point is a disadvantage of the training system of neural network because most researches try to find many approaches to increase the system efficiency to train the data. In contrast, this study excludes the problems of training because the accuracy or error is stemmed from the input data transmitted into the system and we do not know how the data is like. Thus, it is important to let an expert classify each data record by using a method to search for specific characteristics of relationship in the dataset. If any data record has unclear membership and cannot be classified into any cluster, it will be regarded as the noise and the data cleaning must be conducted to remove such record from the dataset.

3 Materials and Methods

3.1 The proposed of Noise filtering Algorithm

A noise filtering of the automatic unsupervised system using the computation intelligence depends on the incremental work which is similar to ILFN [9 10]. Nonetheless, there are certain differences between them. This system can automatically work for initial weight, noise filtering, feature section and data prediction. Also, it supports the dataset and image input. However, this paper includes only the noise filtering which needs to measure the relationship of members in all calculated clusters to find the clarity of relationship. If any record has unclear membership, it will be considered as the noise.

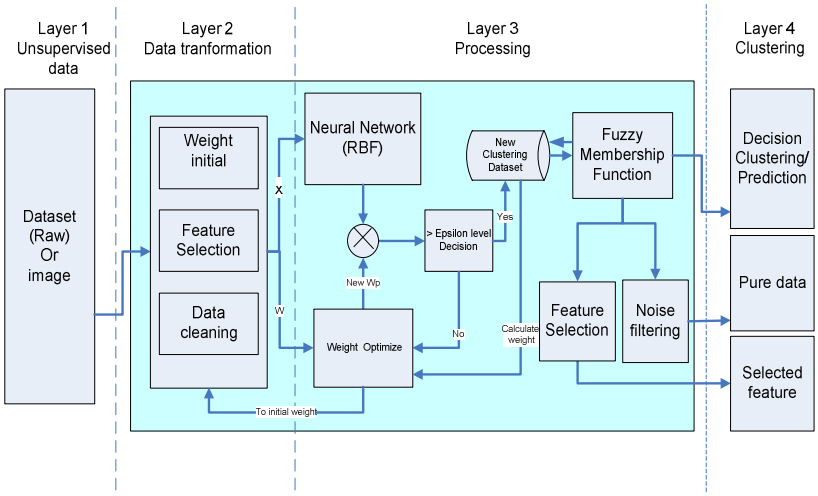

Fig.1. Automatic Unsupervised Clustering using Computational Intelligence Model

The steps of noise filtering are explained as below:

Step 1) Enter input (x) as a vector of the first record into the system.

$$X = [x_1, x_1, \ldots, x_n] \tag{4}$$

Step 2) Measure the distance between the input (x) and the value of weight $(X)^T$ by using Euclidean distant.

Step 3) Define the value of standard division by applying the mean value of the first weight.

Step 4) Calculate the net value which equals y/ standard division.

Step 5) Calculate the e value.

$$e = \left(1 + \left(\frac{1}{\max(net)}\right)\right)^{\max(net)} \tag{5}$$

Step 6) Calculate the membership value (Y) by applying an activate function.

$$Y = \frac{1}{1 + \exp\left(\frac{net^2}{e}\right)} \tag{6}$$

Step 7) Determine the epsilon value = 0.0001 and the centroid of cluster is C = 1.

Step 8) If the membership value > epsilon, adjust the position of weight and standard division in an output class. Then add a new center value (C) and follow Step 2.

Step 9) If the membership value < epsilon, adjust the position of weight and standard division again and follow Step 2.

Step 10) Follow this process to complete the dataset in all records.

3.2 The clarity of the membership

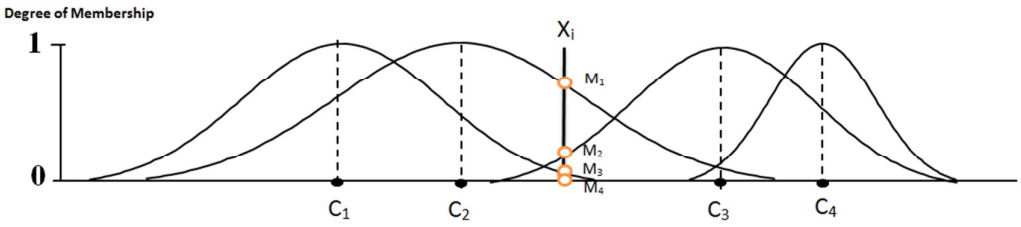

Fig.2. The clarity of membership system.

$$\text{Membership(X,C)} = M_j^{max}(X_i, C_k) \tag{7}$$

Where:

M_j is a degree of record membership of X_i which is compared to the membership of each cluster by using Mahalanobis distance.

X_i is the position of input record.

C_k is the cluster value which contains various values from 1 to k.

According to Figure 2, the position of record of X_i is measured to compare to each members in all clusters (c1, c2, c3, c4) calculated from the unclear membership. The possibility value of membership degree similar to the 2 clusters will be labeled as the noise by conducting a label noise filtering from this equation.

$$\text{NF} = |(member_{\max}(X_i, C_K) - member_{\min}(X_i, C_K))| = 1 \tag{8}$$

Regarding to the above equation, the noise filtering is used to cut the edge of each similar cluster and the difference = 1.

3.3 The performance measurement

To measure the efficiency and accuracy of the noise filtering, this below equation is conducted.

$$\text{ACC}= 100\text{-}\left(\frac{(C_{ans}-C_{noise})}{N} * 100\right) \quad (9)$$

Where

C_{ans} is the number of dataset comprising of the correct answer class obtained from UCI Machine Learning Repository.

C_{noise} is the number of dataset taken by the data cleaning. Nevertheless, some noise records are cut off.

N is the total amount of each dataset.

4 Experimental Results

A simulation of system efficiency test is programmed by applying a tool, Matlab 2010, on intel®Celeron M 1300MHz. This computer consists of 256 KB memory and uses the dataset to test 10 benchmarks including Breast Cancer Wisconsin (Original), Car Evaluation, Contraceptive Method Choice, Glass, Mammographic Mass, Pima Indians Diabetes, Teaching Assistant Evaluation, Acute Inflammations, Vowel Recognition and Wine recognition. These above benchmarks are obtained from Machine Learning Repository of Center for Machine Learning and Intelligent Systems [4] as described below.

1) Dataset of Breast Cancer Wisconsin (Original) is acquired from the hospital in Wisconsin University. The database comprises of 699 sets of 9 types of cancer as follows; clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses. Besides, there are 2 classes of disease – benign and malignant.

2) Dataset of car rental evaluation provides 1,728 samples used for making the decision. Moreover, there are also 6 types of evaluation including buying, maint, doors, persons, lug boot and safety. The results of evaluation contain 4 values – unacc, acc, good and v-good.

3) Database of contraceptive method choice is gathered from a survey in Indonesia. The respondents are 1,473 married and non-pregnant women (including the samples who did not know if they were pregnant during an interview). Several questions are related to wife's age, wife's education, husband's education, number of children ever born, wife's religion, wife's now working, husband's occupation, standard-of-living index and media exposure. Furthermore, there are 3 kinds of contraceptive results such as no-use, long-term and short-term.

4) Dataset of glass consists of 214 records and different features including refractive index, Sodium, Magnesium, Aluminum, Silicon, Potassium, Calcium,

Barium and Iron. These various features are useful to create 7 products, for instance, building windows float processed, building windows non float processed, vehicle windows float processed, vehicle windows non float processed, containers, tableware and headlamps.

5) Database of mammographic mass for today breast cancer is used to predict the probability of breast cancer of 961 samples consisting of 5 types of data; BI-RADS assessment, patient's age in years, mass shape, mass margin and density. The values can be predicted into 2 types – benign and malignant.

6) Database of Pima Indians Diabetes is collected from 768 women over the age of 21. There are 8 features of the data including number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm hg),triceps skin fold thickness (mm),2-hour serum insulin (mu u/ml),body mass index (weight in kg/(height in m)^2),diabetes pedigree function and age (years). An interpretation of disease consists of both positive and negative results as well.

7) Database of teaching assistant evaluation is obtained from an efficiency measurement of instruction during 3 normal semesters and 2 summer semesters. The data is gathered from 151 samples teaching statistics in Wisconsin-Madison University. There are 5 features applied to the evaluation as follows; whether or not the TA is a native English speaker, course instructor, course, summer or regular semester and class size. Besides, 3 kinds of results are indicated – low, medium and high.

8) Database of acute inflammations is acquired from 120 samples. There are 6 structures of this disease including temperature of patient, occurrence of nausea, lumbar pain, urine pushing, micturition pains and burning of urethra. Also, the diagnosis consists of inflammation of urinary bladder and nephritis of renal pelvis origin.

9) Database of vowel recognition contains 528 words and there are total 10 vowel levels of each word. However, there are 11 words pronounced in a real human voice – hid, hId, hEd, hAd, hYd, had, hOd, hod, hUd, hud and hed.

10) Database of wine recognition is used to analyze chemical properties of 178 wines produced from 3 different species in the same area in Italy. There are 13 various chemical properties including alcohol, malic acid, ash, Alkalinity of ash, magnesium, total phenols, flavonoids, non-flavonoid phenols, proanthocyanidins wine, color intensity,hue,od280/od315 of diluted wines and proline. Values that Alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavonoids, non-flavonoid phenols, proanthocyanidins, color intensity, Hue,OD280/OD315 of diluted wines and proline.

Table 1. Number of noises in UCI dataset

	size	#attribute	#class	#noise	% of noise
cancer	699	9	2	2	0.29
car	1728	6	4	322	18.63
cmc	1473	9	3	101	6.86
glass	214	9	6	9	4.21
mamo	961	5	2	44	4.58
pima	768	8	2	132	17.19
tae	151	5	3	45	29.80
urinary	120	6	2	38	31.67
vowel	528	10	11	25	4.73
wine	178	13	3	6	3.37

According to Table 1, there are 10 samples of noise dataset obtained from UCI Machine Learning Repository. Each dataset includes size, number of features of data, number of answer class, number of noise in each dataset and percent of existing noise.

Table 2. Average classification accuracy (%) before noise filtering in UCI data set

	cancer	car	cmc	glass	mamo	pima	tae	urinary	vowel	wine
Logistic	97.00	85.53	52.61	74.30	82.52	78.26	56.29	100.00	77.84	100.00
Multilayer Perceptron	98.86	93.58	58.72	76.17	83.04	80.60	64.24	100.00	91.86	100.00
RBF network	96.42	87.33	48.88	75.23	79.19	74.35	57.62	100.00	95.83	100.00
Simple Logistic	96.71	85.13	52.89	71.03	82.73	78.26	54.30	100.00	77.46	100.00
SMO	97.00	86.40	52.89	61.68	80.12	77.47	54.30	100.00	75.76	99.44
J48	97.28	99.42	74.95	96.26	84.29	84.11	53.64	100.00	95.83	99.44
LMT	96.71	99.48	63.88	99.53	84.39	78.26	58.28	100.00	99.62	100.00
Decision Stump	92.42	70.02	42.70	44.86	81.89	73.57	45.03	59.17	17.61	60.11
Sample Cart	96.71	99.25	59.47	77.10	83.14	77.21	66.89	100.00	95.08	95.51
average	96.57	89.57	56.33	75.13	82.37	78.01	56.73	95.46	80.77	94.94

Table 3. Average classification accuracy (%) after noise filtering in UCI data set

	cancer	car	cmc	glass	mamo	pima	tae	urinary	vowel	wine
Logistic	97.00	86.63	54.38	75.23	83.87	79.17	74.17	100.00	78.22	100.00
Multilayer Perceptron	99.14	96.18	62.39	80.37	83.45	84.38	80.79	100.00	95.08	100.00
RBF network	96.28	94.44	54.24	78.97	79.60	76.04	74.17	100.00	99.05	99.44
Simple Logistic	96.85	87.50	55.26	71.96	83.98	78.13	74.83	100.00	77.08	99.44
SMO	96.85	88.14	54.85	65.42	80.12	78.26	71.52	100.00	75.38	98.88
J48	97.28	99.48	76.37	96.26	84.70	85.55	76.16	100.00	96.21	99.44
LMT	96.85	100.00	62.32	96.73	83.98	78.13	82.78	100.00	96.78	99.44
Decision Stump	92.42	75.87	46.03	48.13	82.52	74.87	62.91	72.50	22.35	72.47
Sample Cart	96.71	99.48	59.81	80.37	83.98	83.46	84.77	100.00	96.97	97.19
average	96.60	91.97	58.41	77.05	82.91	79.77	75.79	96.94	81.90	96.25

Regarding to Table 2 and 3, they present the comparison of data cleaning efficiency gathered from cutting the noise in 10 datasets. The accuracy of data is measured by using classification algorithm in WEKA 3.6 to consider the precision

of each record by comparing to the answer class received from the database of UCI. The 9 algorithms are also used including Logistic, Multilayer Perceptron, RBF network, Simple Logistic, SMO, J48, LMT, Decision Stump and Sample Cart. Furthermore, each algorithm is analyzed and compared before conducting the data cleaning as shown in Table 2. Besides, Table 3 demonstrates the analysis results of algorithm. The dataset with noise filtering is analyzed and compared in mean of each sample group. After data cleaning, the average of dataset with classification algorithm is more precise.

5 Conclusion

According to the noise filtering applying a noise filtering module of the automatic unsupervised clustering using the computational intelligence, it can be noticed that the data with data cleaning is related to the processing following the feature of such data and it is also useful to provide more accurate system analysis without any irrelevant input. Moreover, the comparison of accuracy test, both before and after noise filtering, illustrates that the efficiency of data classification and data analysis increases. Additional, the average of accuracy is better in all algorithms used for the examination.

References

[1] B. V. Dasarathy, Nosing around the neighbor-hood A new system structure and classification rule for recognition in partial exposed environments, Pattern Analysis and Machine Intelligence,1980,67-71.

[2] C. E. Brodley and M. A. Friedl, Identifying and eliminating mislabeled training instances, Proceedings of Thirteenth National Conference on Artificial Intelligence, 1996, 799- 805.

[3] C. E. Brodley and M. A. Friedl, Identifying mislabeled training data, Artificial Intelligence Research, 11(1999),137-167.

[4] Center for Machine Learning and Intelligent Systems, UCI Machine Learning Repository, 2008, accessible via WWW at http://archive.ics.uci.edu/ml.

[5] C. M. Teng, Correcting noisy data, In Proceedings of 16rh international Conference on Machine Learning, 1999, 239-248.

[6] C. M. Teng, Evaluating noise correction, In Proceedings of 6rh Pacific Rim International Conference on Artificial Intelligence, 2000, Lecture Notes in AI, Springer-Verlag.

[7] G. L. Libralon, A. C. P. d. L. F. d. Carvalho and A. C. Lorena,Pre-Processing for Noise Detection in Gene Expression Classification Data, Brazilian Computer Society, 15(2009), 3-11.

[8] G. W. Gates, The reduced nearest neighbor rule, IEEE Transaction on Information Theory, 1972,431-433.
[9] G. Yen and P. Meesad, Constructing a fuzzy expert system using the ILFN network and the genetic algorithm, IEEE International Conference, 2000, 1917-1922.
[10] G. Yen and P. Meesad, Pattern classification by an incremental learning fuzzy neural network, in Proc, IJCNN, 1999**,** 3230-3235.
[11] H. G. Han and J. F. Qiao, Adaptive Computation Algorithm for RBF Neural Network, IEEE Transactions on Neural Network and Learning system, 23(2012), 342-347.
[12] L. Tarassenko and S. Roberts, Supervised and unsupervised learning in radial basis function classifiers, IEEE Proc.-Vis. Image Signal Process, 141(1994), 210-216.
[13] P. Jeatrakul, K. W. Wong and C. C. Fung, Data Cleaning for Classification Using Misclassification Analysis, Advanced Computational Intelligence and Intelligent Informatics, 14(2010), 297-302.
[14] P. E. Hart, The condensed nearest neighbor rule, Institute of Electrical and Electronics Engineers and Transections on Information Theory, 1968, 515-516.
[15] P. Kraipeerapun, C. C. Fung and S. Nakkrasae, Porosity prediction Using Bagging of Complementary Neural Networks, in Advances in Neural Networks, 2009,175-184.
[16] P. Kraipeerapun and C. C. Fung, Binary Classification Using Ensemble Neural Networks and Interval Neutrosophic Sets, Neuro-comput, 72(2009), 2845-2856.
[17] S. Verbaeten and A. Van Assche, Ensemble methods for noise elimination in classification problems, in Multiple Classifier Systems, 2003, 317-325.
[18] X. Zeng and T. Martinez, A Noise Filtering Method Using Neural Networks, International Workshop on Soft Computing Techniques in Instrumentation, Measurement and Related Applications, 2003, 26-34.
[19] X. Zhu, X. Wu and Q. Chen, Eliminating Class Noise in Large Datasets, in Proceedings of the Twentieth Int. Conf. on Machine Learning (20th ICML), Washington D.C. , 2003, 920-927.

Received: July, 2012