Volume 4, Issue 6, June 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Designing Non Assisted Fuzzy Based Model for Text Classification

**Komal**[1]
CSE Department,
Kurukshetra University, India

**Navdeep Kumar**[2]
CSE Department,
Kurukshetra University, India

***Abstract-- Data acquisition and maintenance to any industry is multiplying at an enormous speed. In any industry the produced data is not only massive but also quite complicated. As a result, correct handling of data is of prime importance in order to convert the available data into useful information that leads to knowledge and apposite decision making. Use of data mining in the industry is proving to be a boon for attaining speedy, accurate and futuristic results. Fuzzy Logic is a problem-solving control system methodology that lends itself to implementation in systems ranging from small, embedded micro-controllers to large, multi-channel PC or workstation-based data acquisition and control systems.***

***This paper presents a review of Fuzzy approaches of data mining techniques in current techniques of knowledge discovery. we have present a new algorithm Non-Assisted Fuzzy Based Model for Text Classification for data mining based on Fuzzy Similarity to mine association Rules. The algorithm discussed not only exact matches between items, but also the fuzzy correspondence between them. The algorithm will work unassisted i.e. there should not be a requirement to have an expert for finding similarity between items. It summarizes the techniques to find all association rules that satisfy user-specified minimum support and minimum confidence constraints. Databases using data mining techniques that are in use today.***

***Keywords-- Data Mining, Clustering, Fuzzy, Classification , Non-assisted***

## I. INTRODUCTION

In this paper we present a new algorithm Non-Assisted Fuzzy Based Model for Text Classification for data mining based on Fuzzy Similarity to mine association Rules. The algorithm considers not only exact matches between items, but also the fuzzy similarity between them. The algorithm will work unassisted i.e. there should not be a requirement to have an expert for finding similarity between items. It summarizes the techniques to find all association rules that satisfy user-specified minimum support and minimum confidence constraints. Fuzzy logic is a form of many-valued logic; it deals with reasoning that is approximate rather than fixed and exact. Compared to traditional binary sets (where variables may take on true or false values) fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false.[1]

Fuzzy Logic is a problem-solving control system methodology that lends itself to implementation in systems ranging from simple, small, embedded micro-controllers to large, networked, multi-channel PC or workstation-based data acquisition and control systems. It can be implemented in hardware, software, or a combination of both. FL provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information.

## II. FUZZY CLUSTER ALGORITHM

Let X={$x_1, x_2, \ldots, x_n$} into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centers V, such that

$$\mathbf{V} = v_i | \; i = 1, 2 \ldots c$$

and a partition matrix U such that

$$\mathbf{U = U_{ij}, \; i = 1, \ldots, c, \; j = 1, \ldots, n}$$

Where $U_{ij}$ is a numerical value in [0, 1] that tells the degree to which the element $X_J$ belongs to the i-th cluster.

**Step 1**: Select the number of clusters

$$C, 2 \leq C \leq n,$$

Exponential weight, initial partition matrix $U^0$, and the termination criterion ∈. Also, set the iteration index l to 0.

**Step 2**: Calculate the fuzzy cluster centers

$$\{V_i^1 \; i = 1, 2 \ldots c\}$$

 

**Step 3**: Calculate the new partition matrix $\boldsymbol{U^{i+1}}$

$$\{\boldsymbol{U_i^1}\ \boldsymbol{i} = \mathbf{1}, \mathbf{2} \ldots \boldsymbol{c}\}$$

**Step 4:** Calculate the new partition matrix

$$\Delta = \left|\boldsymbol{U^{i+1}} - \boldsymbol{U^i}\right| = \boldsymbol{max_{ij}}\left|\boldsymbol{U^{i+1}} - \boldsymbol{U^i}\right|$$

**Step 5**: if $\Delta > \in$then, set i = i + 1, and recalculate Fuzzy cluster centers .

**Step 6:** if$\Delta \leq \in$Terminate

The algorithm can be generalized as flows, Fuzzy Clustering involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible. Clustering can also be thought of as a form of data compression, where a large number of samples are converted into a small number of representative prototypes or clusters. Depending on the data and the application, different types of similarity measures may be used to identify classes, where the similarity measure controls how the clusters are formed. Some examples of values that can be used as similarity measures include distance, connectivity, and intensity.

## III. FLOW CHART

Fuzzy Clustering involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible The algorithm start with data scanning , in this items are identified and their domain is find out . for each domain similarity of each item is find out with another item, as similarity is determined for each and every item in a domain. Similar items are then identified in a domain by applying similarity algorithm. Candidates are generated after finding similar items in a domain. Exponential Weight of each candidate is calculated. Evaluation of candidates are done on the basis of weight calculated, such that weight is more or less then the given weights. After evaluation cluster center is identified and all the nearby points that having the distance similarity are grouped together and clusters are formed .Then this clustered output is generated on the basis of similarity.

As shown in figure 1 the flow chart of the process is given. Then weight of the candidate is findout by nearby distance points are evaluated most and clusters centers are identified and then taking nearby points to cluster center a cluster is generated. Cluster output that is generated is based on fuzzy similarity which make itemset in on cluster are more similar to one another and all candidates in one cluster are different from the candidates present in another cluster.

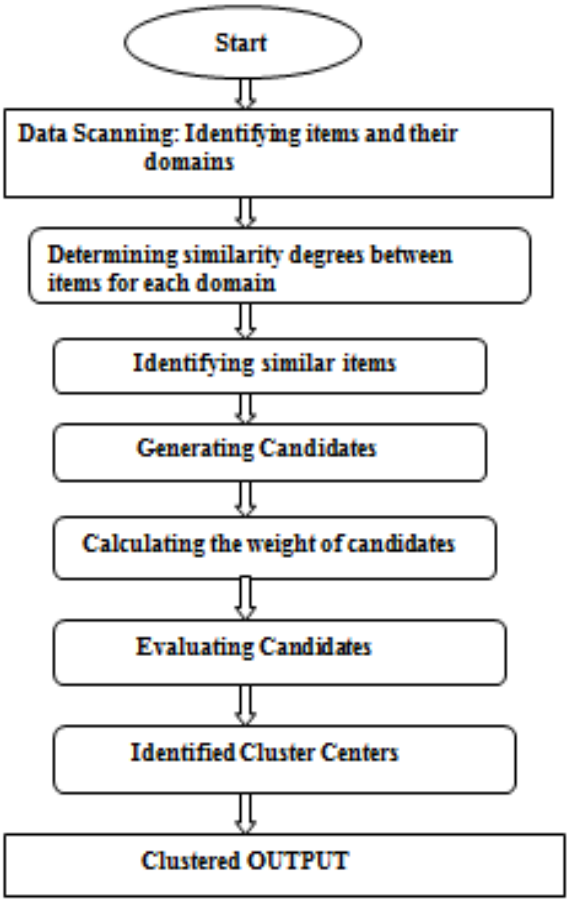


Figure 1 flow chart

 

## IV. RESULTS

### A. Clustering Iris DataSet :

The Iris flower data set or Fisher's Iris data set is a multivariate data set as an example of discriminate analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus". The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. Clustered output on this data set is generated in below figure 2 by using above algorithm with 3 number of cluster as input. Each different symbol represent different cluster center.

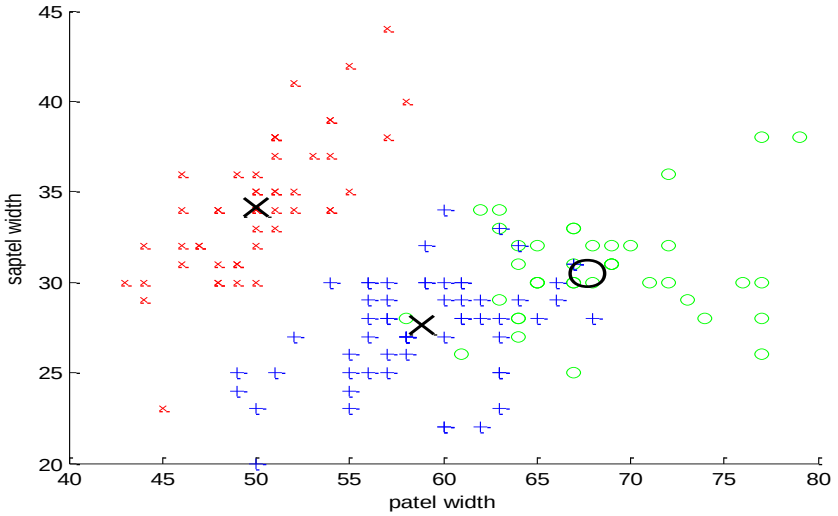


Figure 2: Clustered output

### B. Objective Function

Objective function is also calculated by this algorithm with respect to no. of iteration counts are there. As the graph below represent this relation. Objective function is calculated to find out the performance with respect to iteration count. Line comes in graph near to origin show good performance. As line goes far from origin performance goes down. Performance depend on no. of iteration count occur during process. More the iteration count more accurate is the result.

The goal of the optimization process is to find the parameter values that result in a maximum or minimum of a function called the objective function. Objective function is a mathematical expression describing a relationship of the optimization parameters or the result of an operation (such as simulation) that uses the optimization parameters as inputs as shown in figure 3.

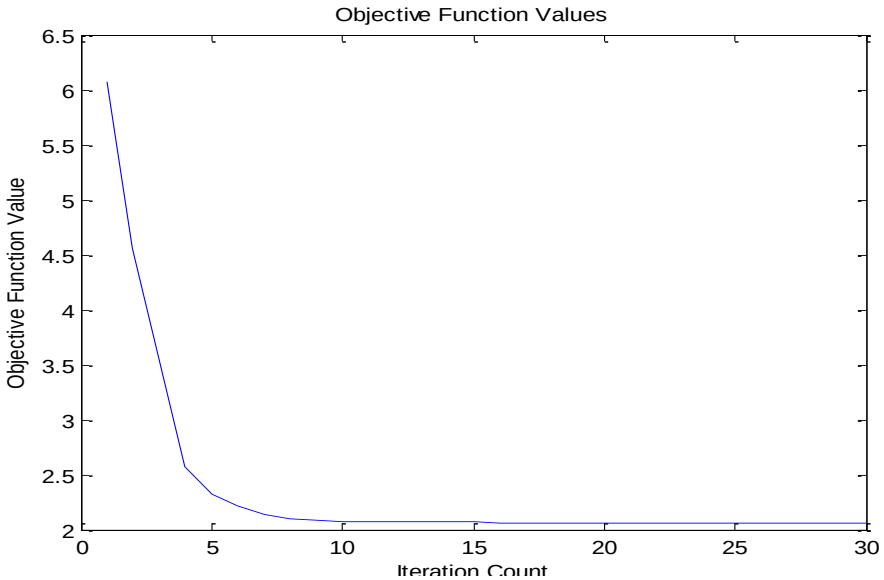


Figure 3: Objective function Graph

 

Objective function value of above examples having 3 or 4 clusters on same data set are compared and result shown on below figure 4. Blue color line in graph is for 4 cluster and red line is for 3 cluster formed on same data set. This graph is obtained by applying the above algorithm. By this graph it is shown that generating 4 cluster for iris data set is better than 3 cluster as 4 cluster take more iteration count than 3 cluster generator. 4 cluster objective function shown in figure 4 below is more near to origin has good performance because of more iteration count is their.

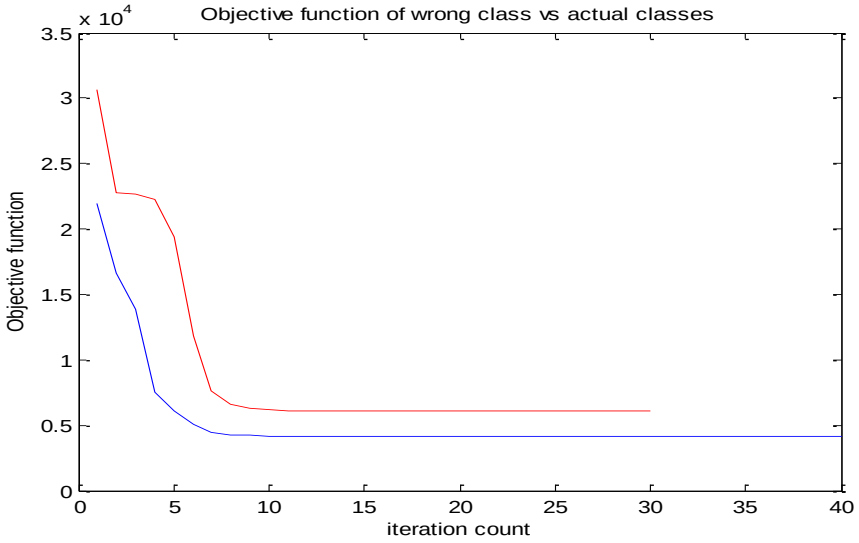


Figure 4: Comparison between objective function value of wrong and actual cluster generated on iris dataset

### C. Result Analysis

The Data set to be optimized given certain constraints and with variables that need to be minimized or maximized using nonlinear Clustering techniques. The fuzzy clustering algorithm can cluster the data set with very good level of accuracy the algorithm is capable of producing fuzzy clusters with good accuracy. The iris data set was taken for evaluation of the algorithm, the data is accurately gets divided into 4 cluster by applying fuzzy clustering algorithm. For good clustering the value of objective function is usually low and the iteration count is reduced significantly. The Figure 4 for accurately shows the difference between correct and incorrect clustering of iris dataset, when accurate no of classes are given the value Objective function of Fuzzy Cluster algorithm is optimal and iteration count is lower than any other class value.

## V. CONCLUSION

The main objectives of the work was to create a Non-Assisted Fuzzy Based Model for Text Classification for data mining based on Fuzzy Similarity. The algorithm considers not only exact matches between items, but also the fuzzy similarity between them. We were able design the data mining algorithm successfully. The Fuzzy clustering algorithm we proposed work unassisted to discover rules and find patterns are more understandable to humans. To meet these challenges, we have propose a new way of obtaining support and confidence for the data mining in such a way that there should not be a requirement to have an expert for finding similarity between items in a given database. This work was an attempt made to summarize all the data points in a dataset using fuzzy data mining algorithm. With the creation and application of data mining, it has become possible to discover clusters rules that reflect the fuzzy similarity among data.

## VI. FUTURE SCOPE

As future work, we want to enhance the computational performance of Fuzzy Clustering algorithm for data mining. We can also plan to define a more refined way of expressing the concepts involved in the data mining in the rules using fuzzy weighted association rules. The Fuzzy clustering algorithm can also be modified to mine time series data like weather clustering and stock portfolio optimization.

### REFERENCES

[1] Novák, V., Perfilieva, I. and Močkoř, J. (1999) *Mathematical principles of fuzzy logic* Dodrecht: Kluwer Academic. ISBN 0-7923-8595-0

[2] Puri, Shalini. "A Fuzzy Similarity Based Concept Mining Model for Text Classification." *International Jtheirnal of Advanced Computer Science & Applications* 2, no. 11 (2011).

[3] Jiang, Jung-Yi, Ren-Jia Liou, and Shie-Jue Lee. "A fuzzy self-constructing feature clustering algorithm for text classification." *Knowledge and Data Engineering, IEEE Transactions on* 23, no. 3 (2011).

[4] Au, W-H., and Keith CC Chan. "Mining fuzzy rules for time series classification." In *Fuzzy Systems, 2013. Proceedings. 2013 IEEE International Conference on*, vol. 1, pp. 239-244. IEEE, 2013.

 

[5] A Padmapriya, KSC Maragatham –" Priority Based Apriori Algorithm For Cancer Prediction Using Fuzzy Classification" International Journal of Engineering …, 2013

[6] OA Tamayo, M Zuluaga, S Ourselin–"Fuzzy classification of brain MRI using a priori knowledge: weighted fuzzy C-means" 2007 – ieeexplore

[7] Lu, S., Hu, H., and Li, F.(2001), "Mining weighted association rules", IntelligentData Analysis 5 (2001),pp.211-225.

[8] Zhang, S., Zhang, C., Yan X., "Post-mining: maintenance of association rules by weighting", Information System 28 (2003) pp.691-707

[9] L.A. Zadeh, "Fuzzy sets," Info. &Ctl., Vol. 8, 2003, pp. 338-353

[10] Daly O, Taniar D. Exception Rules Mining Based on Negative Association Rules, Leture Notes in Computer Science, Vol.3046,2004,pp543-552.

[11] Han. J and M. Kamber (2004), "Data Mining Concepts and Techniques": San Franscisco, CA:.Morgan Kaufmann Publishers.

[12] Chan, K. C. C. and Au, W.-H. (2004) "An Effective Algorithm for Discovering Fuzzy Rules in Relational Databases", In: IEEE International Conference on Fuzzy Systems, Anchorage, Alaska

[13] Ji GL, Yang M. Song YQ. Sun ZH. Fast Updating Maximum Frequent Item sets. Journal of Computers, 2005, 28(1):128-135.

[14] Chen G, Zhu YQ, Yang HB, Study of Some Key Techniques in Mining Association Rule, Journal of Computer Research and Development, 2005, 42 (10): 1785-1789.

[15] T. P. Hong, K. Y. Lin and S. L. Wang, "Mining fuzzy association rules from quantitative transactions", Soft Computing, Vol. 10, No.10, pp. 925-932, 2006.

[16] Gyenesei A, "Mining weighted association rules for fuzzy quantitative items" TUCS Technical Report No. 346, 2006, pp 1-12.

[17] Au, W.-H. and Chan, K. C. C. (2006) "FARM: A Data Mining System for Discovering Fuzzy Association Rules", In: 8th IEEE International Conference on Fuzzy Systems, Seoul, Korea.

[18] David L. Olson, Yanhong Li. "Mining Fuzzy Weighted Association Rules", In: the 40th Hawaii International Conference on System Sciences – 2007 IEEE

[19] Yun, U (2007)., "Efficient mining of weighted interesting patterns with a strong weight and/or support affinity", Information Science 177 (2007) pp.3477-3499.

[20] Preetham Kumar and Ananthanarayana V.S(2008) " Discovery of frequent itemsets using Weighted Tree method" IJCSNS, Vol.8 No. 8 pp. 195-200

[21] Liewean Cheng, Su-Chuan Chen, and Jashen Chen (2009) "Applying Weighted Association Rules with the Consideration of Product Item Relevancy", 978-1-4244-3662-0/09, 2009 IEEE

[22] WeiminOuyang, "Mining Positive Databases and Negative Weighted Fuzzy Association Rules in Large Transaction" (2009) Second International Symposium on Knowledge Acquisition and Modeling pp. 269-272.

 