Improved Swarm Optimization Based C-Means Clustering Technique

Shaveta Saini[1], Rashmi[2]

[1] M.Tech Scholar, Department of Computer Science & Engineerin, Guru Nanak Dev University
RC Jalandhar, Punjab

[2] M.Tech Scholar, Department of Computer Science & Engineerin, Guru Nanak Dev University
RC Jalandhar, Punjab

---***---

Abstract - *The segmentation of the online customized fashion business will be considered as a clustering problem and it is evaluated using various state-of-the art algorithms i.e. Fuzzy C-Means, Naive Baye's (NB) and K-means based mining algorithms. The study indicates that the performance of C-means based mining algorithms for clustering of the online customized fashion business have the better results but it can be improved further by utilisation of the other technique. This paper mainly focuses on the accuracy of clustering rate further by proposing a particle swarm optimization based K-means clustering technique. The particle swarm optimization has ability to find optimistic number of clusters. It will improve the accuracy rate further.*

***Key Words*: Data mining, Clustering, Fuzzy C-means, FCM ,PSO, Particle swarm optimization**

1.INTRODUCTION

Data Mining is defined as getting information and facts via substantial sets of data. To put it differently, most people can say which facts mining is the method with mining knowledge via data. The content or maybe knowledge removed consequently work extremely well in many programs like Market place Evaluation, Deception Diagnosis, Client Storage, Manufacturing Regulate, Scientific research Search etc. Normally, facts mining (sometimes identified as facts or maybe knowledge discovery) is the whole process of examining facts via various aspects along with summarizing the idea straight into valuable information and facts - information and facts which can be used to improve sales revenue, reduces prices, or maybe both. Details mining software is one of the systematic instruments intended for examining data. The idea permits consumers to research facts via several size or maybe angles, sort the idea, along with summarize this interactions identified. Technologically, facts mining is the whole process of obtaining correlations or maybe habits amongst a large number of domains around large relational databases.
Data mining is highly useful in the following domains –

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

1.1 Clustering

Cluster is a group of things, this is probably the same class. In other words, comparable things are classified in a single group as well as different things are classified within one more cluster. This teams a couple of files in a way that enhances your likeness inside of groupings as well as minimizes your likeness in between not one but two different clusters. All these observed groupings can help demonstrate the functions from the root files submitting as well as perform the duties of the cornerstone for alternative files mining as well as analysis techniques.

1.2 Requirement of clustering in Data Mining

- **Scalability** –We want to use hugely scalable clustering algorithms to manage large databases.
- **Ability to deal with different kinds of attributes** – Algorithms needs to be competent to always be put on any type of info just like interval-based (numerical) info, convey, along with binary data.
- **Discovery of clusters with attribute shape** – A clustering algorithm will need to be capable to sensing groups connected with irrelevant shape. Many people should not be surrounded just to length methods that will are inclined to locate rounded bunch connected with small sizes.
- **High dimensionality** – The actual clustering algorithm criteria should not basically equipped to handle low-dimensional facts but the high perspective space.
- **Ability to deal with noisy data** – Databases contain loud, missing or even incorrect data. A few algorithms are generally sensitive to such info as well as can result in low quality clusters.
- **Interpretability** – Clustering final results ought to be interpretable, comprehensible, and usable.

2. FUZZY C-MEANS

Fuzzy C-Mean clustering algorithm is a way to show how data can be classified and clustered in organization or in any application [13]. It was developed by Dunn [14].In this paper, using Fuzzy c-means clustering algorithm background and foreground objects are segmented from the image or frames. This algorithm mainly helps to segment the pixels whether it belongs to background or foreground. The number of clusters is created based on the number of objects in the frames. Applying this fuzzy c means clustering algorithm centroid will be selected. First the centroid is chosen randomly based on the mean of the pixels. The correct centroid will be calculated after finding the degree of pixel using several iterations. In this paper fuzzy c-means clustering method is used for choosing the centroid based on the pixels and the detected edges using the novel edge detection algorithm [11].The following algorithm shows how the fuzzy c-mean clustering technique can be used to segment the foreground object from the given image/frame.

2.1 FCM Algorithm

Step 1: Consider all the pixels at the left vertical axis of the frame.

Step 2: Scan every pixel in horizontal direction until it intersects with the edge pixel which has been detected during edge detection process.

Step 3: Store the value of the edge pixel and continue the process until other edge pixel is found and name those pixels as a, b and so on which gives the boundaries of the object.

Step 4: Continue the scanning process until reach the last pixel in the right vertical axis of the frame.

Step 5: If there is no edge pixel is found then those pixels are considered as background pixels.

Step 6: Calculate the midpoint between a pixel in edge a and a pixel in edge b using the formula

$$\text{Midpoint}\ (a, b) = \frac{(x_a, y_a) + (x_b, y_b)}{2}$$

Where (xa, ya) is the coordinate of a pixel in edge a and (xb, yb) is the coordinate of a pixel in edge b.

Step 7: Identify the mid points for all edge pixels by applying the Midpoint formula.

Step 8: Connect all the mid points column wise.

Step 9: Locate the column which contains more number of midpoints and mark the first and last midpoint in that column.

Step10: Calculate the midpoint for that column using the midpoint.

3. PARTICLE SWARM OPTIMIZATION

The Particle swarm optimization (PSO) is a technique to get search engine marketing in metaphor associated with social conduct associated with flocks associated with parrots and/or institutions associated with fish. A lot like inherited algorithms (GAs), the actual PSO can be the optimizer based on population. The machine is definitely initialized to begin with throughout a few with little thought created possibilities methods, and then is conducted to locate the actual greatest just one iteratively. The technique may be developed by way of a sim associated with easy social models. PSO draws on swarms for instance sea food instruction and chicken flocking. Using the investigation latest shopping results for chicken preferring, parrots have found meals by simply preferring (not by simply just about every individual). Like GA, PSO have to furthermore have a health evaluation function that may the actual particle's location and assigns with it some sort of health value. The task using the greatest health importance inside total run is termed the international most effective (P,). Each particle additionally monitors the top health value. The positioning with this importance is termed the private most effective (e). The basic formula consists of sending your line some sort of populace associated with debris in the research room, remembering the most beneficial (most fit) answer encountered. At intervals of time, every particle adapts the pace vector, based on the momentum as well as affect associated with both the most practical answer as well as most practical answer of that others who live nearby, and then computes a different denote examine.

3.RELATED WORK

F. Provost et al. (2013) [48] proposed that Organizations possess came to the realization you have to seek the services of facts people, academic corporations ended up struggling to set up data-science products, and also publications ended up offering facts technology to be a popular employment choice. Nonetheless, there was some sort of distress concerning what precisely facts technology seemed to be, knowing that distress may lead to disillusionment while the notion diffuses directly into incomprehensible buzz. Here, there are described several explanations why that it was not easy to flag down accurately precisely what is facts science.**F.M. Hsu et al. (2012)** proposed the segmentation of buyers is crucial for a corporation desperate to grow ideal campaign approaches for different clusters. Clustering buyers is an in-depth comprehension of their behaviour. However, former researchers have given little focus the likeness of countless pieces of transaction. Deficiency of different types along with thought amounts of objects, comes from item-based segmentation strategies are certainly not as good as expected. By making use of a thought power structure of things, this research states a segmentation method to spot commonalities amongst customers.[44]**R.S. Wu et aussi al.**

(2011) suggested the particular segmentation concerning on line people in multiple types may help with a much better realizing in addition to depiction involving behaviour inside the electronic digital marketing market. Online shopping directories include things like multiple sorts of data upon buyer acquiring pastime in addition to market qualities, and also use characteristics for example World wide web utilization in addition to full satisfaction by using services. Information about buyers exposed by way of segmentation allows firm facilitators to build good buyer operations in addition to perfect their own promotion tactics to enhance buyer expectations. To accomplish optimal segmentation, most of us created smooth clustering strategy that works with a hidden mixed-class account clustering strategy to label on line buyers dependent on their acquiring files over categories. A method based on the particular hidden Dirichlet allocation product is actually accustomed to create the buyer segments. Variance approximation is actually leveraged to build estimations from your segmentation in a computationally-efficient manner. A consist of smooth clustering strategy assure far more appealing benefits in comparison with challenging clustering in addition to more significant within-segment clustering good quality compared to the specific fusion product.[43]**F. Herrera et al. (2011)** proposed Subgroup uncovering that's a information exploration technique which often extracts helpful regulations with regards to a focus on variable. A crucial characteristic of this can be lots of people connected with predictive plus detailed induction. A review related towards undertaking connected with subgroup uncovering can be presented. This kind of examine focuses on the foundations, algorithms, plus sophisticated scientific studies plus the applications of subgroup uncovering provided through the specific bibliography.[40]**H.D. Park et al. (2009)** proposed a brand new algorithm formula to get K-medoids clustering which often rans including the K-means algorithm formula as well as subjected to testing many strategies for deciding on first medoids. This suggested algorithm formula assessed the space matrix after as well as completed it to find completely new medoids at most iterative step. To be able to measure the suggested algorithm formula, the item was applied with a few true as well as man made info units as well as weighed against the outcome connected with additional algorithms the modified Rand index. Trial and error outcomes demonstrate that the actual suggested algorithm formula had a considerably reduced period in working out by using very similar performance from the partitioning all-around medoids.[46]

4. GAPS IN LITERATURE

Following are the various gaps in earlier work.

1. The use of hybridization of data mining techniques can be done to improve the accuracy rate further for clustering of the online customized fashion business.
2. The integration of k-means and particle swarm optimization has been ignored to improve the accuracy rate further for clustering of the online customized fashion business.
3. The accuracy rate of the existing methods is found to be poor so improvement is required to make them more consistent.

5. METHODOLOGY AND RESULTS

5.1 Methodology

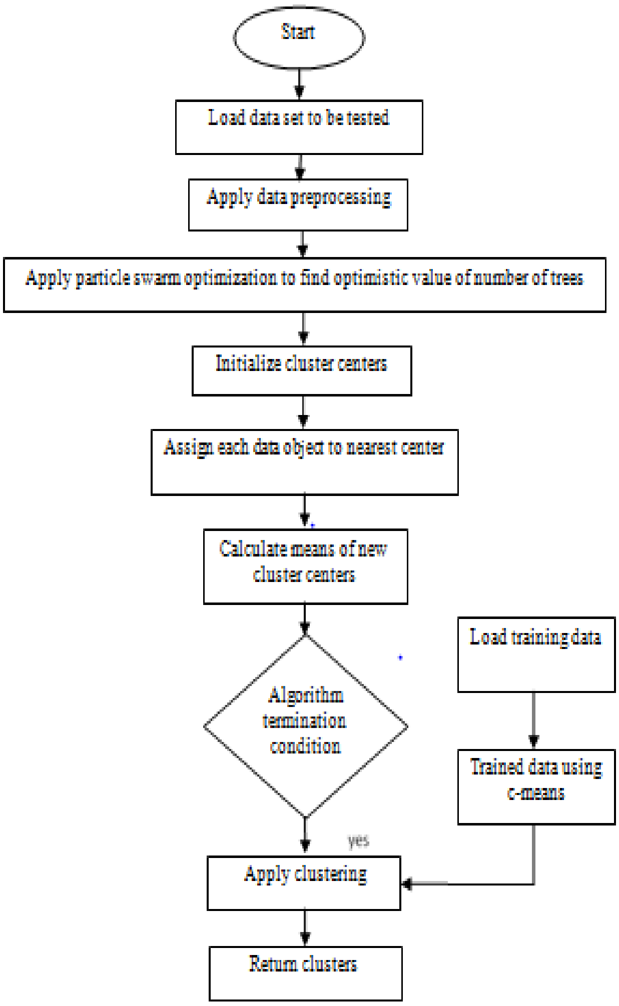
Start
Load data set to be tested
Apply data preprocessing
Apply particle swarm optimization to find optimistic value of number of trees
Initialize cluster centers
Assign each data object to nearest center
Calculate means of new cluster centers
Algorithm termination condition
Load training data
Trained data using c-means
yes
Apply clustering
Return clusters

Figure1.Proposed methodology

5.2 Performance Analysis

This paper has designed and implemented the proposed technique in MATLAB tool u2013a. The evaluation of proposed technique is done on the basis of following metrics i.e. Accuracy, F-measure, true positive rate and false positive raate. A comparison is drawn between all the parameters with proposed algorithm and figures shows all the results.

1. Accuracy

Accuracy refers to the ability of the model to correctly predict the class label of new or unseen data.
It is calculated as-

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where

True positives (TP) =No. of correct classifications predicted as yes(or positive).
True negatives (TN) =No. of correct classifications predicted as no(or negative).
False positive (FP) =No. of incorrect classifications predicted as yes(positive) when it is actually no(negative).
False negative (FN) =No. of incorrect classifications predicted as no(negative) when it is actually yes(positive).

2. F-Measure-It is the measure that combines precision and recall. It is the harmonic mean of precision and recall. It is calculated as-

$$F = 2 * \left[\frac{(Precision * Recall)}{Precision + Recall}\right]$$

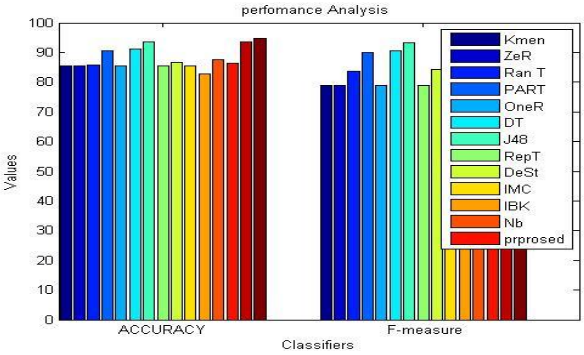

Figure2. Analysis of accuracy and F-measure

3. Precision- Precision is a description of random errors, a measure of statistical variability.

$$Precision = \frac{TP}{TP + FP}$$

4. Recall- Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance

$$Recall = \frac{TP}{TP + FN}$$

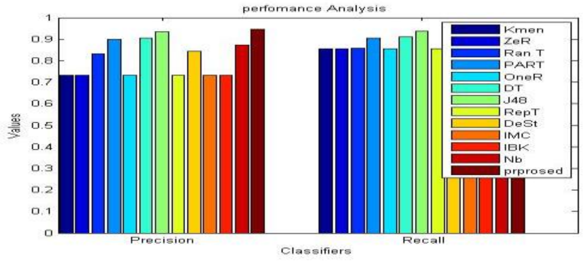

Figure3. Analysis of precision and recall

5.True positive rate- TPR measures the proportion of positives that are correctly identified as such (i.e. the percentage of sick people who are correctly identified as having the condition).

$$TPR = \frac{TP}{TP + FN}$$

6.False positive rate- FPR usually refers to the probability of falsely rejecting the null hypothesis for a particular test. The false positive rate is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events

$$FPR = \frac{FP}{FP + TN}$$

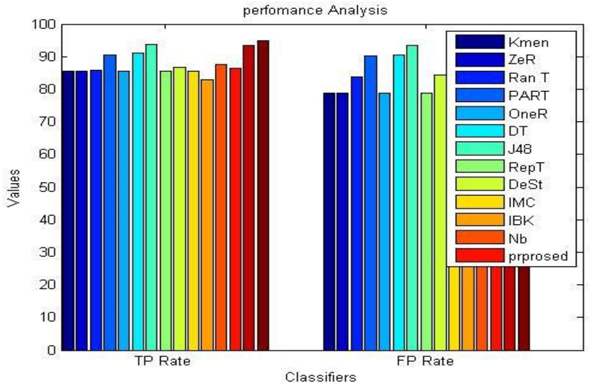

Figure4. Analysis of TPR and FPR

6.CONCLUSION

In this paper, we have analyzed existing the performance of C-means based mining algorithms for clustering of the online customized fashion business. The proposed particle swarm optimization based C-means clustering technique gives the better results. This paper has shown comparison between exiting and proposed work on the basis of parameters like True Positive Rate, False Positive Rate, Accuracy, F-measure, Precision and Recall. By comparing the existing and proposed technique the qualitative analysis has clearly shown that the main improvement in proposed work. But in proposed work PSO do not guarantee an optimal solution is ever found. Also, PSO does not use the gradient of the problem being optimized So in near future we will try to enhance the results further by using the differential evolution technique so the more improvement can be done.

REFERENCES

[1] Yiming Huang, Di Wu, Yinshui He, Na Lv, Shanben Chen, "The selection of arc spectral line of interest based on improved K-medoids algorithm",Advanced Robotics and its Social Impacts (ARSO) 2016 IEEE Workshop on, pp. 106-109, 2016, ISSN 2162-7576.

[2] Kennedy, J., 2011. Particle swarm optimization. In Encyclopedia of machine learning (pp. 760-766). Springer US.

[3] Masson, R., Iosif, L., MacKerron, G. and Fernie, J., 2007. Managing complexity in agile global fashion industry supply chains. The International Journal of Logistics Management, 18(2), pp.238-254

[4] Vandenbroucke, B. and Kruth, J.P., 2007. Selective laser melting of biocompatible metals for rapid manufacturing of medical parts. Rapid Prototyping Journal, 13(4), pp.196-203.

[5] Poli, R., Kennedy, J. and Blackwell, T., 2007. Particle swarm optimization. Swarm intelligence, 1(1), pp.33-57.

[6] Atzmueller, M., & Puppe, F. (2006, September). SD-Map–A fast algorithm for exhaustive subgroup discovery. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 6-17). Springer Berlin Heidelberg.

[7] Pal, N.R., Pal, K., Keller, J.M. and Bezdek, J.C., 2005. A possibilistic fuzzy c-means clustering algorithm. IEEE transactions on fuzzy systems, 13(4), pp.517-530.

[8] Lavrač, Nada, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. "Subgroup discovery with CN2-SD." Journal of Machine Learning Research 5, no. Feb (2004): 153-188.

[9] Agarwal, Nitin, Ehtesham Haque, Huan Liu, and Lance Parsons. "Research paper recommender systems: A subspace clustering approach." In International Conference on Web-Age Information Management, pp. 475-491. Springer Berlin Heidelberg, 2005.

[10] Christopher, Martin, Robert Lowson, and Helen Peck. "Creating agile supply chains in the fashion industry." International Journal of Retail & Distribution Management 32.8 (2004): 367-376.

[11] Da Silveira, G., Borenstein, D., & Fogliatto, F. S. (2001). Mass customization: Literature review and research directions. International journal of production economics, 72(1), 1-13.

[12] Wagstaff, Kiri, Claire Cardie, Seth Rogers, and Stefan Schrödl. "Constrained k-means clustering with background knowledge." In ICML, vol. 1, pp. 577-584. 2001.

[13] Ronen, Simcha, and Oded Shenkar. "Clustering countries on attitudinal dimensions: A review and synthesis." Academy of management Review (1985): 435-454.

[14] [14] Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241-266.

[15] Raikwal, J. S., & Saxena, K. (2012). Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. *International Journal of Computer Applications*, *50*(14).

[16] Moses, D. (2015). A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data. *Kuwait Journal of Science*, *42*(2).

[17] Nichat, A. M., & Ladhake, S. A. (2016). Brain Tumor Segmentation and Classification Using Modified FCM and SVM Classifier. *Brain*, *5*(4).

[18] Verma, L., Srivastava, S., & Negi, P. C. (2016). A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. *Journal of Medical Systems*, *40*(7), 1-7.

[19] Li, D. C., Liu, C. W., & Hu, S. C. (2010). A learning method for the class imbalance problem with medical data sets. *Computers in biology and medicine*, *40*(5), 509-518.

[20] Kazemzadeh, R. S., & Sartipi, K. (2005, September). Interoperability of data and knowledge in distributed health care systems. In *13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)* (pp. 230-240). IEEE.