ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Hierarchical Clustering Algorithm for Improved Incomplete Pattern Classification

Priyanka G Harankhedkar, Dr. A. N. Banubakode

Student, Department of Computer Engineering, RSCOE, Savitribai Phule Pune University, Pune, India

HOD, Department of Information Technology, RSCOE, Savitribai Phule Pune University, Pune, India

ABSTRACT: while working with the database it is very often that the values are not given properly or might me missing for some fields, but those are important and must be known for further working. It can be happen due to the person entering the values might not know the exact value or failure of sensors or may be leave the empty space by mistake. The categorization of missing valued incomplete pattern is a hard task in machine learning. The Incomplete data is not useful in classification process. Using prototype values when incomplete patterns are classified, the final class for the same patterns possibly produced inaccurate output. We cannot assign particular class for particular pattern. In such cases system can produce the faulty results which can outcomes in diverse effects. So for sorting this issue of incomplete data, we propose to implement prototype-based credeal classification (PCC) method. This technique is fused with Hierarchical clustering and Evidential reasoning strategy to give accurate, time and memory efficient outcomes. PCC can trains the samples and find out the class prototype. In this way this might be helpful for recognizing the missing values. After having all the unknown values, creedal technique can be used for classification. The test result shows the updated version of PCC works better in terms of time and memory efficiency..

KEYWORDS: Belief functions, hierarchical clustering, credal classification, evidential reasoning, missing data.

I. INTRODUCTION

Data mining can be treated as a strategy to discover legitimate information from huge datasets and identifying patterns. Such examples are further helpful for classification process. The fundamental use of the data mining procedure is to discover helpful information inside dataset and deliver it into a proficient arrangement for future use.

In a large portion of the classification issue, some attribute fields of the item are not filled. There are different reason for the empty attributes including disappointment of sensors, incorrect values field by user, sometime didn't get the meaning of field so client leave that field vacant and so on. There is a need to discover the efficient technique to classify the item which has missing attribute values. Different classification techniques are there, in literature, to manage the classification of efficient pattern. Some of the methods deletes the missing valued patterns and just uses complete patterns for the classification process. In any case, at some point half patterns contain essential data consequently, so this technique is not a proper answer. Additionally this strategy is relevant just when deficient information is under 5% of entire information. Leaving out the fragmented information might diminish the quality and execution of classification algorithm. Next strategy is just to fill the missing values. It is likewise time consuming procedure. This paper depends on the classification of incomplete patterns. On the off chance that the missing values relate a lot of information features then evacuation of the information feature might come about into a more prominent loss of the required appropriate data. So this paper essentially focuses on the classification of incomplete patterns.

Hierarchical clustering creates a cluster hierarchy or a tree-sub tree structure. Each cluster node has descendants. Ordinary clusters are blended or spilt as indicated by the top-down or bottom up methodology. This technique helps in finding of information at various levels of tree.

At the point when incomplete patterns are classified utilizing model values, the last class for the same pattern might have various outcomes that are variable outcomes, with the goal that we can't characterize particular class for particular examples. While computing prototype value utilizing normal computation might don't have enough memory and time in results. To conquer this problem, proposed framework actualizes evidential capabilities to compute particular class

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

for particular pattern and hierarchical clustering to calculate the model, which yields effective results as far as time and memory.

A. *Motivation*:

Information might be incomplete. This study proposes a classification for fragmented information. Classifying the information with missing values is a repetitive task and the outcome of the procedure relies on the system used to handle the fragmented information. Different methods have been developed before, having their own particular pros and cons identified with classifying incomplete patterns. So there is a need of distinguishing the efficient method to illuminate issues identified with incomplete pattern classification. To pick the right strategy for taking care of this viewpoint one must have appropriate knowledge of the different factor that outcome in the missing information. Efficient treatment of missing values requires a complete comprehension behind it.

II. RELATED WORK

In paper [2], by conducting several experiments, author come to an end that, generally, imputation increases the subsequent classification, other than for the mean imputation method which provides improvements only when used for a substantial (50%) amount of missing data.

In paper [3], author developed a unique technique to makes use of ensemble networks to handle the missing data issue. The thought is to produce the missing values more than once to make a few clean datasets. These datasets are utilized to prepare a few neural networks to make an ensemble. Authors give three variations of the strategy that contrast as indicated by the way we create the missing values. Authors also introduced an analytical analysis that shows the effectiveness of the proposed strategies.

In paper [4], authors have studied the impact of the imputation methods in two admiration measures. These two measures are the Wilson's noise ratio and the average mutual information difference. The first one quantifies the noise induced by the imputation technique in the instances which contain MVs. The second one analyzes the increment or decrement in the relationship of the isolated input attributes concerning the class label. Authors have observed that the CMC and EC strategies are the ones which gives less commotion and keep up the shared data better, which responds to the best imputation methods observed for each FRBCS types.

In paper [5], author studied missing information imputation procedures with the point of developing more precise algorithm. Author carried the idea from fuzzy K-means clustering, and apply it to the issue of missing information imputation. The test results exhibit the strength of this technique. They also evaluate the performance of the algorithms in light of the RMSE error analysis. We find that the fundamental K-implies algorithm beats the mean substitution technique, which is a common and regular methodology for missing information imputation.

In paper [6], Author talk about the way of the combinations (conjunctive versus disjunctive, revision versus updating, static versus dynamic data fusion), contend about the requirement for a normalization, look at the conceivable beginnings of the conflicts, figure out whether a combination is supported and examine several numbers of the proposed arrangements.

In paper [7], Author proposes one way to deal with predict membership or belief functions from information. The strategy depends on the support vector regression of membership or belief functions. In this way, they have changed the current support vector regression approach with a specific end goal to combine the same requirements of the membership and belief functions. Authors can do either straight or nonlinear regression utilizing or not a kernel. Authors additionally propose an answer for improvement issue utilizing the SMO approach.

III. FINDINGS

In above papers, a ton of new procedures are concentrated on for the classification of the incomplete patterns. The procedures utilize significantly the estimation of the missing values and after then classification. Existing framework

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

utilizes Mean Imputation (MI) procedure for calculating prototypes in system. While computing prototype value utilizing average calculation and clustering algorithms might prompts wasteful memory and time in results.

IV. PROBLEM STATEMENT

To solve time, memory and erroneous result issues, proposed framework executes evidential thinking to compute particular class or Meta class for particular pattern and hierarchical clustering to compute the prototype, which yields proficient results as far as time and memory.

V. PROPOSED ALGORITHM

A System Architecture

In this framework, we are developing another procedure to classify the intense or almost difficult to sort information with the assistance of conviction function Bel(.).In our proposed framework we are preparing our framework to take a shot at missing information from dataset. As an input we are utilizing fragmented example dataset as information For this execution. For development we can utilize any standard dataset with missing fields. Currently available framework was utilizing Mean Imputation (MI) procedure for computing prototype in framework. We are utilizing K-Means clustering as a first piece of our development. K-Means clustering gives additional time and memory proficient results for our framework than that of mean imputation (MI) system.

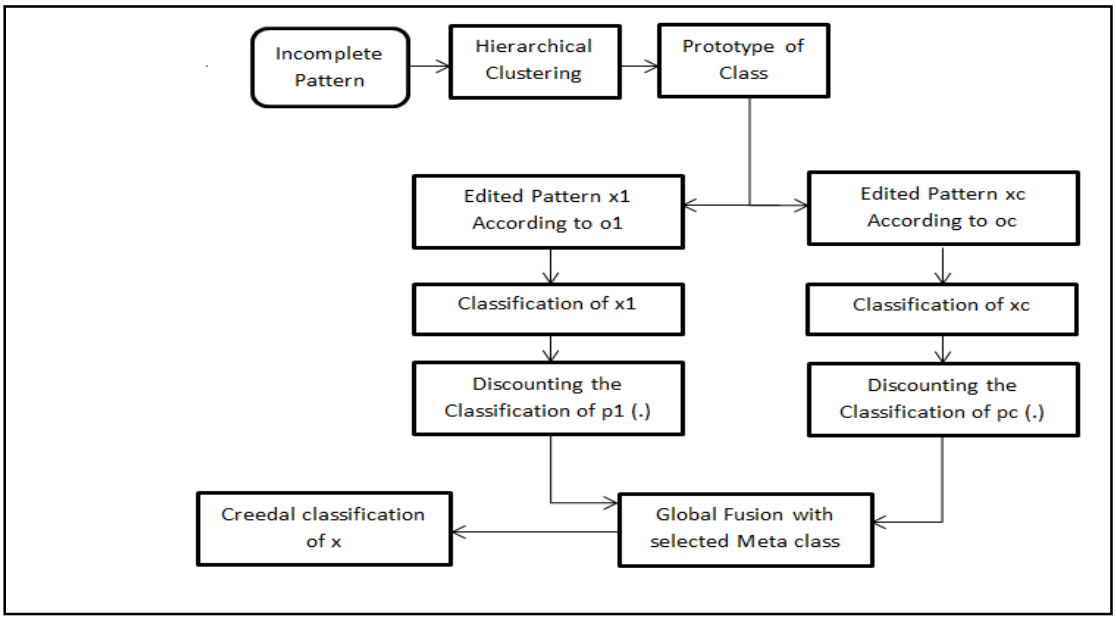

Fig.1. System Architecture

Second part of our proposed framework is to utilize hierarchical clustering for prototype computation. Hierarchical clustering gives more efficient results as contrast with that of K-Means clustering. Consequently we are concentrating on particularly progressive grouping which is utilized at purpose of model creation.

After Prototype development, we are utilizing the KNN Classifier to characterize the examples with the models computation set up of the missing qualities. Since the separation between the object and the prototype model is diverse we are utilizing the discounting technique for the characterization. We then fuse the classes by utilizing the global fusion rule and then as indicated by the threshold value. Threshold value gives the number of the object that should be incorporated into the Meta classes. In this way we expand the accuracy by mishitting the object into particular class if there should be a case of the uncertainty to classify in one class.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

After that we can apply special method to categories the objects into one specific class. In proposed framework we are concentrating on time efficiency while formation of the prototype.

B. Algorithms

Algorithm 1 Hierarchical Algorithm:

Input: P objects from dataset

Method:-

1: Amongst the input vector points calculate a distance matrix
2: Every data point must be considered as a cluster.
3: Repeat step 2

4: Combine two nearly similar clusters.

5: Alter distance matrix

6: Go to step 3until the single cluster remains

7: Stop

Output: Clusters of similar vector.

C. Mathematical Model

System M is represented as M= {S, P, E }

- Initial Phase
  Input Incomplete Pattern
  S= {s1, s2, s3, …..,sn}
  Where, S is the set of objects with missing values

- Process

1. Prototype Calculation
   P={p1,p2,…pn}
   Where P is the set of prototypes

2. Classification Results

$$P_i^g = \Gamma(x_i^g \mid Y)$$

Where $\Gamma(.)$ represent the chosen classifier, and is the output of the classifier.

3. Calculation of reliability factor

$$\alpha_i^g = \frac{W_i^g}{W_i^{\max}}$$

Where,

$$W_i^{\max} = \max(w_i^1, \ldots\ldots w_i^c)$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

4. Calculating Discounting mass of Belief

$$\begin{cases} m_i^g(A) = \alpha_i^g P_i^g(A), & A \subset \Omega \\ m_i^g(\Omega) = 1 - \alpha_i^g + \alpha_i^g P_i^g(\Omega) \end{cases}$$

5. The set of potential classes

$\hat{}i=\{W_s \mid Bel(W_{max})\text{-}Bel(W_s) < \varepsilon\}$

Where $\varepsilon \in [0,1]$ is chosen threshold.

- Output:
  E= { e1, e2, e3, ….en }

Where, E is the final class of the object. Either single or the meta classes is obtained

D. Experimental Setup

For building the system we used Java framework (version jdk 8) on windows system. As a development tool we are using Net beans (version 8.1). The system will be work on any standard machine. In case of experiments we are writing code for missing data from dataset.

VI. RESULTS AND DISCUSSION

A. Dataset

Dataset used for proposed system is Yeast Data Set that is of Protein Localization Sites which includes 8attributes and 10 classes with total size of 93 KB. This dataset is collected from UCI Machine Learning Repository (i.e. https://archive.ics.uci.Edu/ml/datasets/Yeast). Only 10 to 20 % data or values will be missing in case of the incomplete patterns.

B. Result Set

The outcome set for the paper is for the most depending on the time and memory comparison of the old and the new proposed framework architecture.

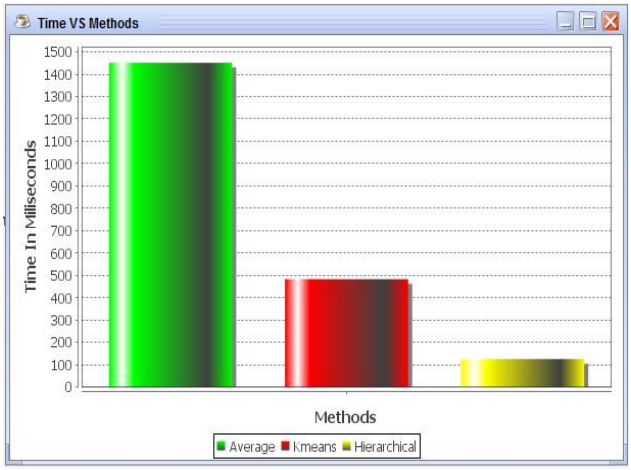

Fig. 3. Time comparison graph

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

From graph we can see time consumption of the old system and proposed system. As we can see that proposed system takes very less time to compare with the old or existing system. Proposed system takes minimum time because it uses hierarchical clustering algorithm for prototype calculation and classification of edited patterns.

Hierarchical clustering algorithm is more efficient than K- means algorithm.

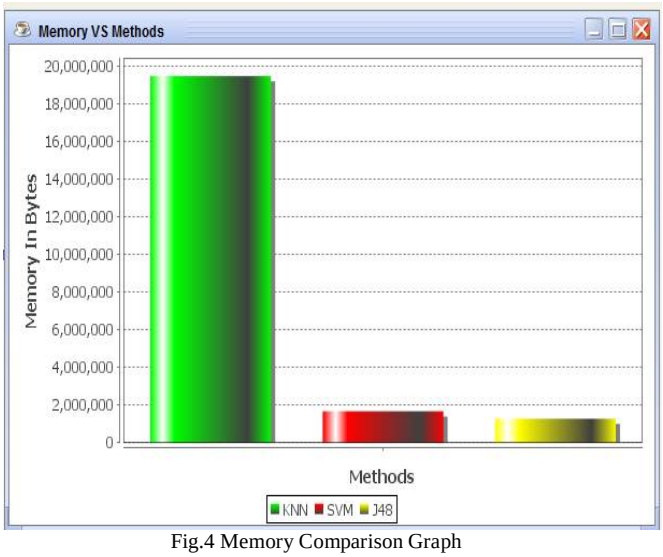

Fig.4 Memory Comparison Graph

Graph demonstrates the memory use by existing framework and proposed framework. As should be obvi that proposed framework uses less memory as contrast and the old or existing framework

C. Algorithmic Complexity

The algorithmic complexity will be the number of iterations that are required to classify an in-complete pattern object properly to the specific class.

VII. CONCLUSION AND FUTURE SCOPE

Proposed a missing example classification for fragmented object operation that calculates a value and pattern by arithmetic formula belief function. In proposed technique evidential reasoning shows vital role for missing patterns in the dataset. After the discounting strategy utilizing the belief function and the threshold of the Meta classes the objects with incomplete pattern is classified. In the event that most results square measure dependable on a classification, the article will be focused on a selected category that is effectively committed to the most well-known result.However, the high clash between these outcomes infers that the classification of the article is kind of uncertain or inexact solely supported the far-famed attributes data. In such case, the article turns out to be horribly difficult to classifications properly in an exceedingly specific class and it's reasonably designated to the privilege meta-class laid out by the mix of the exact classes that the article is likely be having a place.At that point the conflicting mass of belief is allotted completely to the chosen meta-class. On the off chance that the incomplete pattern object is dispensed to a meta-class, it shows that the exact classifications enclosed inside of the meta-class seem indistinguishable for this object supported the far-famed attributes.

This system will be improved in following ways:

- From manual observation user can give prototype value.
- For implemented hierarchical clustering algorithm to calculate the prototype value various clustering algorithm can be re-placed.
- To classify final class from meta-classes new methods can be used.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

REFERENCES

1. Zhun-Ga Liu, Quan Pan, Grgoire Mercier, and Jean Dezert, "A New Incomplete Pattern Classication Method Based on Evidential Reasoning", North-westernPolytechnical Uni-versity, Xian 710072, China,4, APRIL 2015
2. Farhangfar, Alireza, Lukasz Kurgan, and Jennifer Dy. "Impact of imputation of missing values on classification error for discrete data." Pattern Recognition 41.12 (2008): 3692-3705.
3. Hassan, Mostafa M., et al. "Novel ensemble techniques for regression with missing data." New Mathematics and Natural Computation 5.03 (2009): 635-652.
4. Luengo, Julián, José A. Sáez, and Francisco Herrera. "Missing data imputation for fuzzy rule-based classification systems." Soft Computing 16.5 (2012): 863-881.
5. Li, Dan, et al. "Towards missing data imputation: a study of fuzzy k-means clustering method." Rough sets and current trends in computing. Springer Berlin Heidelberg, 2004.
6. P. Smets, Philippe. "Analyzing the combination of conflicting belief functions." Information fusion 8.4 (2007): 387-412.
7. H. Laanaya, A. Martin, D. Aboutajdine, and A. Khenchaf, "Support vector regression of membership functions and belief functions—Application for pattern recognition," Inform. Fusion, vol. 11, no. 4,pp. 338–350, 2010.
8. G. Batista and M. C. Monard, "A study of K-nearest neighbour as an imputation method," in Proc. 2nd Int. Conf. Hybrid Intell. Syst., 2002, pp. 251260.
9. O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, no. 6, pp. 520525, 2001.
10. G. Batista and M. C. Monard, "A study of K-nearest neighbour as an imputation method," in Proc. 2nd Int. Conf. Hybrid Intell. Syst., 2002, pp. 251260.

.