International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 25 No. 1 Jun. 2016, pp. 320-331

http://www.ijisr.issr-journals.org/

HIGH DIMENSIONAL K^N-FAST CLUSTERING BASED INTELLIGENT DECISIVE SUPPORT SYSTEM FOR EFFICIENT DISEASE PREDICTION USING DATA MINING AND RULE SETS

D. BANUMATHY[1] and S. SELVARAJAN[2]

[1]Assistant Professor / CSE, Paavai Engineering College, Namakkal, Tamilnadu, India

[2]Principal, Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

ABSTRACT: The methods of high-dimensional clustering have been applied for variety of problems and in case of decisive support systems, there are few approaches discussed earlier, but suffers with the problem of false indexing ratio with poor clustering accuracy and higher time complexity. To overcome the issue of poor clustering accuracy, a novel K^n Fast Clustering algorithm is discussed in this paper. The method generates rule sets using the data records from the data set. First the dimension N is identified and for each dimension the range values are identified. From identified fuzzy values, the method computes disease impact factor for each of the dimension or symptoms towards each disease class. Based on the impact factor and the data points, we generate rule sets that consist of a single rule for each of the disease class. The K^n Fast clustering algorithm uses the fuzzy rule sets generated and for each data point from the data set, the clustering algorithm computes K^N dimensional similarity measure. Based on computed similarity measure, the data points are assigned a class, and the method reduces the false indexing, overlapping, and time complexity of clustering.

KEYWORDS: High-Dimensional Clustering, Decisive Support System, Disease Prediction, Data Mining, Rule Sets.

1 INTRODUCTION

The problem of clustering high dimensional data set has been adapted in many situations. Clustering data points with a small size can be done in an easier way by computing similarity measure between the data points of a different class of data points. The problem becomes tougher when the size of dimension grows and computing the similarity measure between data point also becomes difficult. This introduces a false indexing ratio, and the same data point may be assigned with multiple class names, where the similarity between data points has to be computed by considering all the dimensional values.

The high-dimensional clustering can be applied in a variety of ways and can be adapted to the problem of decision making. The decisive support system is one which provides support to the clinical process, and there are many decisive support systems which works based on symptoms provided. The symptoms based decisive support systems, generates recommendations on different diseases according to input symptoms given. Similarly, the decisive support systems can use the high dimensional data set to perform such recommendations.

The input symptoms set or medical diagnosis results can be used to perform disease prediction. Based on the given input symptoms or diagnosis results, the probability or inference about any disease can be performed. For example, for a general fever the symptoms may be of body pain, temperature and cold but differs with the symptoms of Typhoid. For small dimensional symptoms or diagnosis results the prediction of disease can be performed easily but when the size of dimension grows, then the data points has to be clustered in proper manner so that the prediction of disease can be done in an efficient manner.

Data mining is the process of extracting useful and required information from the large knowledge base. There may be a large amount of information present in the knowledge base and to produce an efficient result to the support system, the data retrieval, and indexing has to be done in an efficient way. The disease prediction system may require the similar records to be extracted in short time, to produce such result the method has to produce efficient cluster. The data points can be clustered using some rules, and the rules can be generated using the values of the data point. For each data point, there may be N number of attributes and each has different values. By computing range values of each attribute, the rules can be generated. From the rule set, the data points can be clustered using some similarity measure.

K^N Fast clustering is one, which cluster the data points according to the N-dimensional similarity of data points and attributes. By computing the overall similarity of data attributes between any two data point can be computed and based on the similarity measure they can be grouped into one class. This approach increases the speed of clustering and reduces the time complexity as well.

2 RELATED WORKS

The researchers have described a number of approaches for the problem of decisive support systems and this section discuss some of the methods of decisive support system.

Projected Clustering with LASSO for High Dimensional Data Analysis [1], uses attribute selection and handles the sparse structure of the data effectively. We select the most informative attributes that do preserve cluster structure using LASSO (Least Absolute Selection and Shrinkage Operator). Though there are other methods for attribute selection, LASSO has distinctive properties that it selects the most correlated set of attributes of the data. This model also identifies dominant attributes of each cluster which retain their predictive power as well. The quality of the projected clusters formed is also assured with the use of LASSO.

Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence [2], has been developed to cluster data using high-dimensional similarity based PCM (SPCM), with ant colony optimization intelligence which is effective in clustering nonspatial data without getting knowledge about cluster number from the user. The PCM becomes similarity based by using mountain method with it. Though this is efficient clustering, it is checked for optimization using ant colony algorithm with swarm intelligence. Thus, the scalable clustering technique is obtained, and the evaluation results are checked with synthetic datasets.

Towards Unsupervised and Consistent High Dimensional Data Clustering [3], a modified PROCLUS algorithm is proposed, aimed at improving the running time and consistency as well as the unsupervised selection of the parameter like, average number of dimensions. The promising and consistent results of MPROCLUS has open the sky wide open for further research for usage of MPROCLUS in Stream Data Mining.

Hubness-Based Clustering of High-Dimensional Data [4], review and refine existing work which explains the mechanisms of the phenomenon, establishes the location of hub points near central regions of clusters in the data, and shows how hub ness can negatively affect existing clustering algorithms by virtue of hub points lowering between-cluster distance. Next, we review the newly proposed partition clustering algorithms, based on K-means, which take advantage of hub ness by employing hubs in the process of cluster prototype selection. These “soft” K-means extensions avoid premature convergence to suboptimal stable cluster configurations and can reach the global optima more often. The algorithms offer significant improvements over the K-means baseline in scenarios involving high-dimensional and noisy data. The improvements stem from a better placement of hub points into clusters, which helps in increasing the between-cluster distance. Finally, we introduce novel clustering algorithms as “kernelized” versions of the most successful hub ness-based methods discussed above, that can more effectively handle arbitrarily shaped clusters.

In Model-based clustering of high-dimensional binary data [5], a mixture of latent trait models with common slope parameters for model-based clustering of high-dimensional binary data, a data type for which few established methods exist, is proposed. Recent work on clustering of binary data, based on a d-dimensional Gaussian latent variable, is extended by incorporating common factor analyzers. Accordingly, this approach facilitates a low-dimensional visual representation of the clusters. The model is further extended by the incorporation of random block effects. The dependencies in each block are taken into account through block-specific parameters that are considered to be random variables. A variational approximation to the likelihood is exploited to derive a fast algorithm for determining the model parameters. Real and simulated data are used to demonstrate this approach.

Clustering high throughput biological data with B-MST, a minimum spanning tree based heuristic [6], address important challenges in bioinformatics, high throughput data technologies are needed to interpret biological data efficiently and

reliably. Clustering is widely used as a first step to interpreting high dimensional biological data, such as the gene expression data measured by microarrays. A good clustering algorithm should be efficient, reliable, and effective, as demonstrated by its capability of determining biologically relevant clusters. This paper proposes a new minimum spanning tree based heuristic B-MST, that is guided by an innovative objective function: the tightness and separation index (TSI). The TSI presented here obtains biologically meaningful clusters, making use of co-expression network topology, and this paper develops a local search procedure to minimize the TSI value. The proposed B-MST is tested by comparing results to: (1) adjusted Rand index (ARI), for microarray data, sets with known object classes, and (2) gene ontology (GO) annotations for data sets without documented object classes.

A Relevant Clustering Algorithm for High- Dimensional Data [7], is used for finding the subset of features. A Relevant clustering algorithm renders efficiency and effectiveness to find the subset of features. Relevant clustering algorithm work can be done in three steps. First step elimination of irrelevant features from the dataset; the relevant features are selected by the features having the value greater than the predefined threshold. In the second step selected relevant features are used to generate the graph, divide the features using graph theoretic method, and then clusters are formed by using Minimum Spanning Tree. In the third step find the subsets features that are more related to the target class is selected.

Fast agglomerative clustering using information of k-nearest neighbors [8], develop a method to lower the computational complexity of the pairwise nearest neighbor (PNN) algorithm. Our approach determines a set of candidate clusters being updated after each cluster merge. If the updating process is required for some of these clusters, k-nearest neighbors are found for them. The number of distance calculations for our method is O(N2), where N is the number of data points. To further reduce the computational complexity of the proposed algorithm, some available fast search approaches are used. Compared to available approaches, our proposed algorithm can reduce the computing time and number of distance calculations significantly. Compared to FPNN, our method can reduce the computing time by a factor of about 26.8 for the data set from a real image. Compared with PMLFPNN, our approach can reduce the computing time by a factor of about 3.8 for the same data set.

Appling the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis [9], applied dimensionality reduction and automated clustering methods on human T helper (TH) cells derived from peripheral blood and tonsils, which showed differential cell composition and extensive TH cell heterogeneity. Notably, this analysis revealed numerous subtypes of follicular helper T (TFH) cells that followed a continuum spanning both blood and tonsils. Furthermore, we identified tonsillar CXCR5loPD-1loCCR7lo TFH cells expressing interferon-γ (IFN-γ), interleukin-17 (IL-17), or Foxp3, indicating that TFH cells exhibit diverse functional capacities within extrafollicular stages. Regression analysis demonstrated that CXCR5loPD-1− and CXCR5loPD-1lo cells accumulate during childhood in secondary lymphoid organs, supporting previous findings that these subsets represent memory TFH cells. This study provides an in-depth comparison of human blood and tonsillar TFH cells and outlines a general approach for subset discovery and hypothesizing of cellular progressions.

Clustering High-Dimensional Landmark-based Two-dimensional Shape Data [10], develop a penalized model-based clustering framework to cluster landmark-based planar shape data, while explicitly addressing these challenges. Specifically, a mixture of offset-normal shape factor analyzers (MOSFA) is proposed with mixing proportions defined through a regression model (e.g., logistic) and an offset-normal shape distribution in each component for data in the curved shape space. A latent factor analysis model is introduced to model explicitly the complex spatial correlation. A penalized likelihood approach with both adaptive pairwise fusion Lasso penalty function and L2 penalty function is used to realize automatically variable selection via thresholding and deliver a sparse solution.

Robust estimation of the mean vector for high-dimensional data set using robust clustering [11], a robust starting point for S-estimator based on robust clustering is proposed which could be used for estimating the mean vector of the high-dimensional data. The performance of the proposed estimator in the presence of outliers is studied and the results indicate that the proposed estimator performs precisely and much better than some of the existing robust estimators for high-dimensional data.

Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence [12], has been developed to cluster data using high-dimensional similarity based PCM (SPCM), with ant colony optimization intelligence which is effective in clustering non-spatial data without getting knowledge about cluster number from the user. The PCM becomes similarity based by using mountain method with it. Though this is efficient clustering, it is checked for optimization using ant colony algorithm with swarm intelligence. Thus, the scalable clustering technique is obtained, and the evaluation results are checked with synthetic datasets.

Hubness-based fuzzy measures for high-dimensional knearest neighbor classification [13], propose several fuzzy measures for k-nearest neighbor classification, all based on hubness, which express fuzziness of elements appearing in k-neighborhoods of other points. In Timeseries classification in many intrinsic dimensions [14], demonstrated that it can be sufficient to induce hubness: a phenomenon where some points in a data set participate in unexpectedly many k-nearest neighbor lists of other points. After explaining the origins of hubness and its interaction with the information provided by labels, we formulated a framework which, based on hubness and the distribution of label mismatches within a data set, categorizes time-series data sets in a way that allows one to assess whether hubness can be used to improve the performance of the k-NN classifier.

Random Projection Towards the Baire Metric for High-Dimensional Clustering [15], use random projection with the following principle. With the greater probability of close-to-orthogonal projections, compared to orthogonal projections, we can use rank order sensitivity of projected values. The Baire metric, divisive hierarchical clustering, is of linear computation time.

A graph based subset selection approach has been discussed in [16], to cluster high dimensional data. The method identifies the most impact features of each class based on which the subset of features is formed. The method describes that the features belong to any class are independent. The evaluation results show that the method has a higher probability in identifying the subset and produces high-quality results.

All the above-discussed approaches have the problem of false indexing and overlapping that reduces the clustering efficiency.

3 K^N FAST CLUSTERING-BASED DISEASE PREDICTION

The K^n clustering approach groups the data points into a number of groups according to the similarity measure computed. The algorithm computes the similarity measure in each dimension of data points at each class. Finally, the method computes the cumulative similarity measure of input data point with the rest of the data points of each class. Based on computed similarity measure the method chooses a class to which the input data point belongs to.

The proposed K^N clustering approach for decisive support system has a number of stages namely Rule generation, K^N Clustering, and Disease Prediction. We explain each of the functional components in detail in this section.

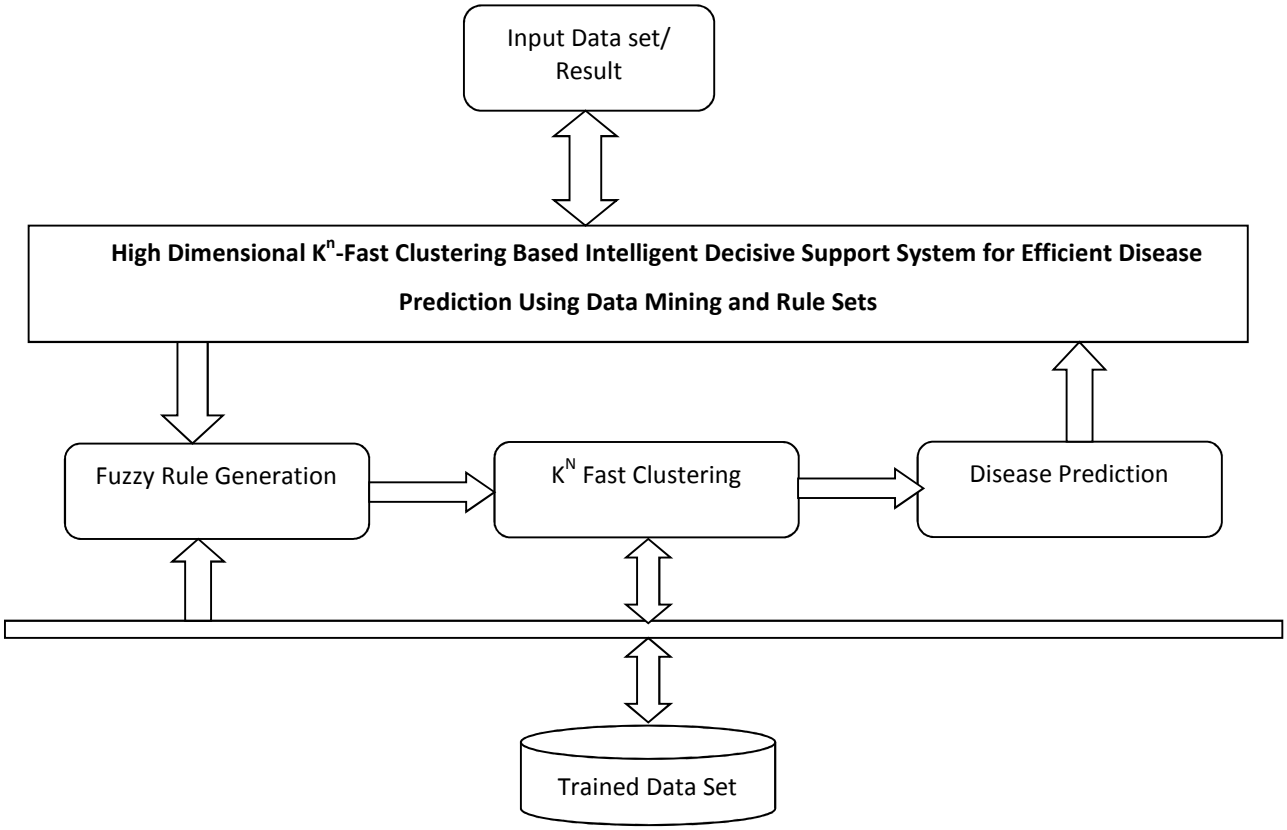


Figure 1: Proposed System Architecture

The Figure 1 shows the architecture diagram of the proposed high-dimensional clustering algorithm for disease prediction. Also, it shows the functional components of the proposed approach.

3.1 FUZZY RULE GENERATION:

The fuzzy rule generation is performed as the preprocessing stage of the proposed approach. The method first identifies the dimension of the data set and then the method computes the minimum and maximum values of each of the attribute identified from the data set. With the range values identified, the method computes the disease impact factor for each dimension to the different classes of disease. Based on the computed DIF value, the attribute selection is performed to generate a rule. Using the selected attributes and minimum and maximum values, the method generates the fuzzy rule for each disease class.

Algorithm:

Input: Data Set Ds

Output: Rule Set Rs.

Start

Identify Attribute Set As = $\sum_{i=1}^{size(Ds)} \sum Attr(Ds(i)) \nexists As$

For each attribute Ai from As

Compute Maximum value Amax = $\int_{i=1}^{size(Ds)} Ds(i).Ai > Amax$

Compute Minimum value Amin = $\int_{i=1}^{size(Ds)} Ds(i).Ai > Amin$

for each disease Di

compute disease impact factor DIF.

DIF = $\frac{\sum_{i=1}^{size(Ds)} \sum Ds(i).Ai > Amin \cup Ds(i).Ai < Amax \;\&\&\; Ds(i).Disease == Di}{\sum_{i=1}^{size(Ds)} \sum Ds(i).Ai > Amin \cup Ds(i).Ai < Amax}$

End

if DIF > DTh //DTH- Disease Threshold

Add Attribute and values to disease Attribute set.

DAset = $\sum Ak(DAset) \cup Ai$

End

End

for each disease Di

generate rule Ri.

Ri = $\bigcup \sum Ai(DAset), Amin, Amax$

Add to rule set Rs = $\sum Rk(Rs) \cup Ri$

End

Stop

The above-discussed algorithm generates the fuzzy rule for each disease using which the data set can be clustered in an efficient manner.

3.2 K^N FAST CLUSTERING

The clustering algorithm computes multi dimensional similarity measure with each of the data points. For each data point, the method computes the multi-dimensional (K^N) disease impact closure Similarity with the rule present in the rule set. The closeness of each dimension represent the similarity of any two data point in one dimension, and this will be repeated for all dimensions to compute the similarity in all the dimensions. The computed multi-dimensional disease impact closeness similarity shows the closeness of any two data point in overall. Based on computed MDDICS value, the method assigns a class to the data point.

Algorithm:

Input: Data Set Ds

Output: Cluster Cs

Start

for each data point Di from Ds

for each disease class Ci from Cs

compute MDDICS.

$$\text{MDDICS} = \sum_{k=1}^{\text{size(Rs)}} \frac{\sqrt{\Sigma(Ri.A1-Ci(Ri).A1)2+\cdots.+(Ri.An-Ci(Ri).An)2}}{\text{size(Rs)}}$$

end

choose less distanced class Ci = Min(MDDICS).

Assign Di with class Ci.

End

Stop

The above-discussed algorithm performs clustering of data points, by computing multi-dimensional closure similarity measure. Based on computed similarity the method assigns labels to the data points and index them into the clusters.

3.3 DISEASE PREDICTION

The disease prediction is performed by computing the disease impact similarity measure between the data points and the rule set available. For each rule available in the rule set, the method computes the multi-dimensional disease impact similarity measure. Using the similarity measure computed, the method identifies the disease class that has more disease impact similarity and the disease is concluded as the more probable one.

4 RESULTS AND DISCUSSION

The proposed K^N fast clustering algorithm based decisive support system has been implemented using Matlab with various data sets. The approach has been validated for its clustering efficiency using various data sets. The method has been tested with a training set of 80 percent data, and the remaining is used as test set. The proposed method produced efficient results with various data sets. For the evaluation purpose, the popular Wisconsin data set has been used, which consists of varying attributes and values. The data set has ten dimensions and in overall the data set has 569 instances.

We have evaluated the algorithm with the following data sets.

TABLE 1: Description of Data Sets

Dataset	Number of Data Points (N)	Attributes (d)	Attribute Values (AA)	Classes (K)
Wisconsin	569	10	Multi Variant	2

The table 1 shows the list of data sets has been used to evaluate the performance of the different algorithm.

The Wisconsin data set Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Separating plane described above was obtained using Multi-surface Method-Tree (MSM-T) [22], a classification method that uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes. The data set has two classes namely malignant and benign.

For the evaluation purpose, different test cases have been carried out. For example, a symptom set with the following attributes and values are given as follows:

Table 2: Details of input value for the test case

Attribute Name	Range value	Original Value Submit
Sample code number: id number	x	2
Clump Thickness	1-10	7
Uniformity of Cell Size	1-10	6
Uniformity of Cell Shape	1-10	4
Marginal Adhesion	1-10	10
Single Epithelial Cell Size	1-10	3
Bare Nuclei	1-10	6
Bland Chromatin	1-10	5
Normal Nucleoli	1-10	4
Mitoses	1-10	1

The Table 2, shows the test case value given for the evaluation of the proposed method. It shows the values of attributes given for the input test sample. The data set has two classes of data points where the value 2 represents the benign class and the value 4 represents the malignant class.

The method generates fuzzy rule sets from the input data set given at the training phase using all the dimensions. The rule has a range of values for each dimension for any specific class. For example, the rule generated for the class Benign would be something like displayed in Table 3.

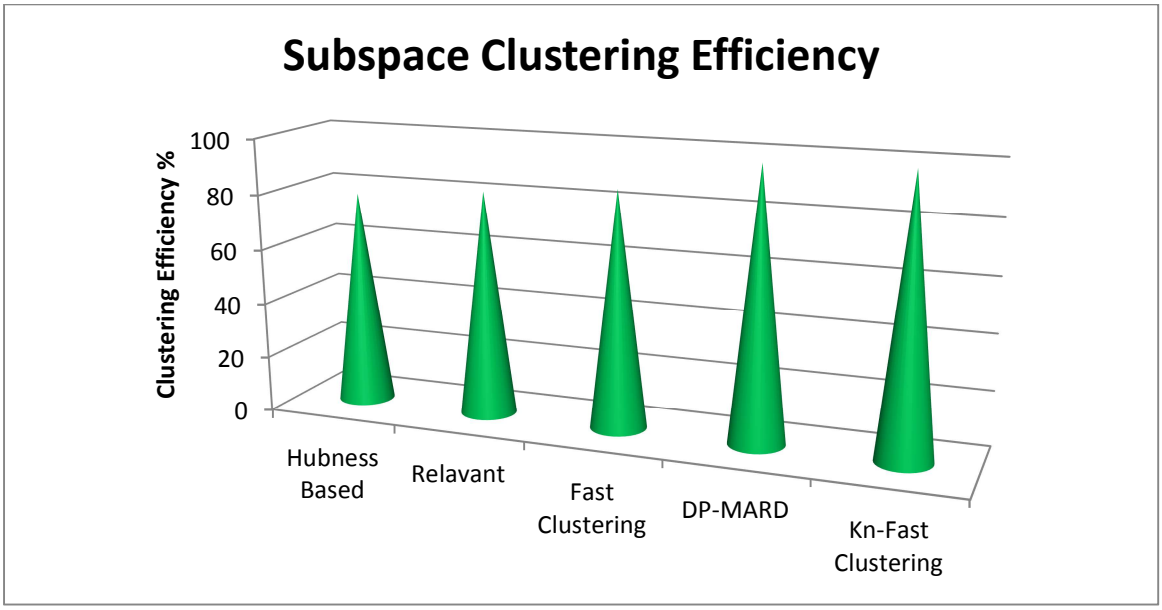
Table 3: Generated Rule Sets for both the classes benign and malignant

Attribute	Range	Class
Sample code number: id number	Any	Malignant
Clump Thickness	7-10	
Uniformity of Cell Size	6-10	
Uniformity of Cell Shape	7-10	
Marginal Adhesion	7-10	
Single Epithelial Cell Size	8-10	
Bare Nuclei	5-10	
Bland Chromatin	6-10	
Normal Nucleoli	8-10	
Mitoses	7-10	
Sample code number: id number	Any	Benign
Clump Thickness	1-7	
Uniformity of Cell Size	1-6	
Uniformity of Cell Shape	1-7	
Marginal Adhesion	1-7	
Single Epithelial Cell Size	1-8	
Bare Nuclei	1-5	
Bland Chromatin	1-6	
Normal Nucleoli	1-8	

The Table 3 shows the rule being generated for the class benign and malignant. It is visible that the both the cases have different range of values for each dimensional attribute.

Based on this information, the input symptoms set or attribute values can be used to compute the multi-Dimensional Disease Impact Closure Similarity. When the input value is mapped with the rule set being generated and computes the MDDICS value to perform the classification of the input data point. From the Table 2, the similarity values of the input data point has more than 60 percent similarity to the class malignant, because most of the values of the attributes lies within the

range of the rule of malignant class. Similarly, the test has been carried out with a large number of data points in different categories.



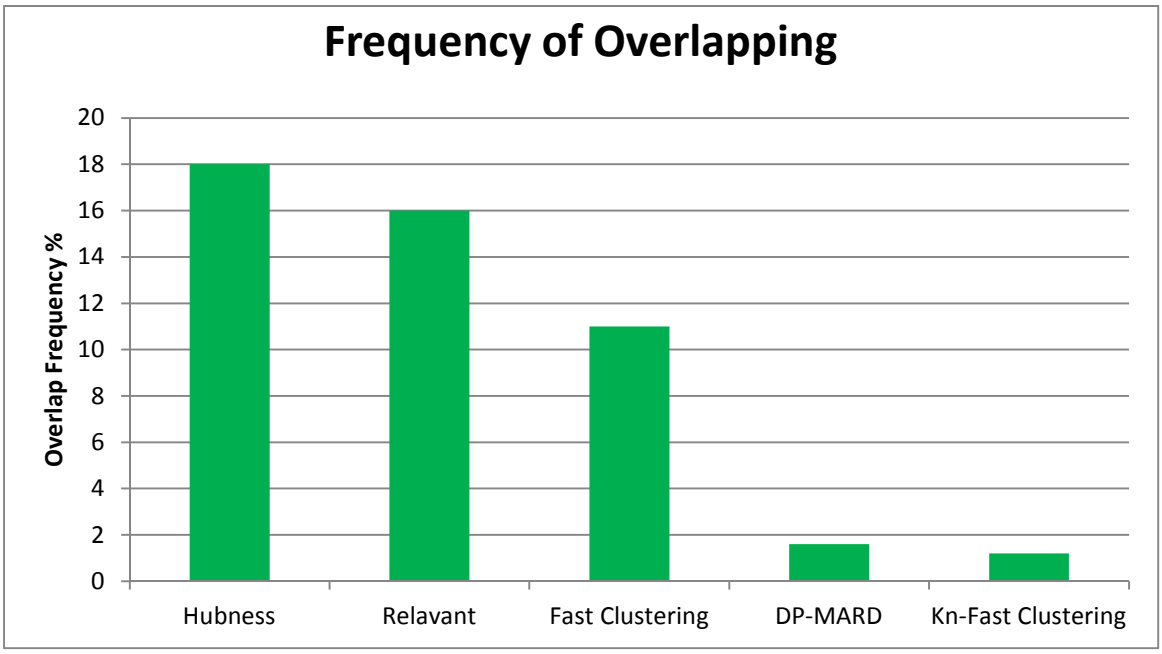
Graph 1: Comparison of clustering efficiency

The Graph1 shows the comparison of clustering efficiency produced by different methods; it shows clearly that the proposed method has produced more accurate cluster than others.

Table 4: Comparison of overlapping and false indexing of different methods

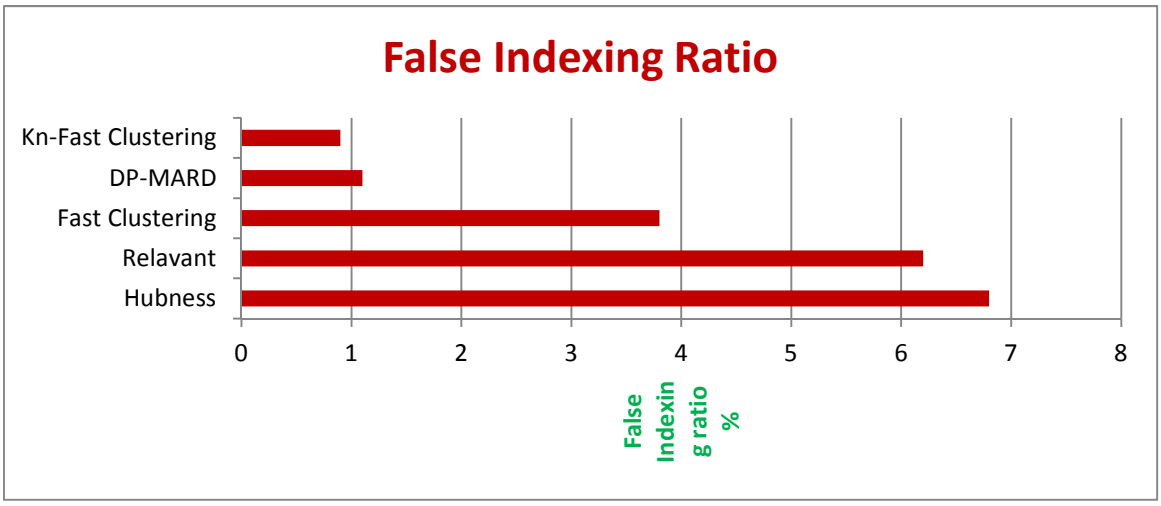
Method Name	Overlapping Ratio	False Indexing Ratio
Hubness	18	6.8
Relavant	16	6.2
Fast Clustering	11	3.8
DP-MARD	1.6	1.1
Kn-Fast Clustering	1.2	0.9

The Table 4 shows the ratio of overlapping and false indexing produced by different methods. The values show that the proposed method has produced less ration in both the factors.



Graph2 shows the data point overlap frequency of different algorithm

The Graph2 shows the comparison of overlap frequency produced by different methods. It shows clearly that the proposed method has produced less overlap than other methods that will be removed at the next stage.



Graph3: Shows false indexing ratio of different algorithms

The graph3 shows the ratio of false indexing produced by different algorithms for the same size of training and testing set. It shows clearly that the proposed approach has produced efficient results with least false indexing ratio.

Table 5: Comparative study of result

Data set	Hubness	Relevant	Fast Clustering	DP-MARD	Kn-Fast Clustering
Wisconsin at 70/30 Test case	78.67	81.52	86.67	97.8	98.6
Wisconsin at 90/10 Test case	80.76	84.12	88.97	99.63	99.72

The Table 5 presents the comparative study of results produced by different methods and it shows clearly that the proposed method has produced a more efficient result than other methods. The results show that the K^N Clustering technique has produced more efficient results in all the factors of disease prediction and has been verified with the various data sets considered.

5 CONCLUSION

We proposed K^N fast clustering based disease prediction algorithm using fuzzy rule sets. The method generates the rule set from the input data set given at the training phase. For each disease class, the method generates rule by computing min-max values for each attribute of the data point. The attributes are selected according to the value of min-max and which has greater than the threshold. Using the identified attribute and values, the method generates the rule set. Using the rule set the method computes multi-dimensional disease impact similarity measure to cluster the data points. The same measure is used to predict the disease and produces efficient results with more accuracy and less time complexity.

REFERENCES

[1] Lidiya Narayanan, Anoop S. Babu, M. R. Kaimal, Projected Clustering with LASSO for High Dimensional Data Analysis, Springer, Advances in Intelligent Systems and Computing Volume 327, 2015, pp 201-209.

[2] Thenmozhi Srinivasan and Balasubramanie Palanisamy, Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence, Hindawi, the scientific world journal, 2015.

[3] R G Mehta, N J Mistry and M Raghuwanshi. Article: Towards Unsupervised and Consistent High Dimensional Data Clustering. International Journal of Computer Applications 87(2):40-44, February 2014.

[4] Nenad Tomašev, Miloš Radovanović, Dunja Mladenić, Mirjana Ivanović, Hubness-Based Clustering of High-Dimensional Data, Springer, Partitional Clustering algorithms, pp:353-386, 2015.

[5] Yang Tang, Ryan P. Browne and Paul D. McNicholas, Model based clustering of high-dimensional binary data, Computational Statistics & Data Analysis, 2015, vol. 87, issue C, pages 84-101.

[6] Pirim H, Ekşioğlu B, Perkins AD, Clustering high throughput biological data with B-MST, a minimum spanning tree based heuristic, Comput Biol Med. 2015 Jul 1;62:94-102.

[7] Bini Tofflin.R1 , A Relevant Clustering Algorithm for High- Dimensional Data , International Journal of Innovative Research in Computer and Communication Engineering Vol.2, Special Issue 1, March 2014

[8] C.-T. Chang, J. Z. C. Lai, and M. D. Jeng, "Fast agglomerative clustering using information of k-nearest neighbors," Pattern Recognition, vol. 43, no. 12, pp. 3958–3968, 2010.

[9] Michael T. Wong3, Jinmiao Chen3, Sriram Narayanan, Wenyu Lin, Rosslyn Anicete, Henry Tan Kun Kiaang, apping the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis,Cell Reports, Volume 11, Issue 11, p1822–1833, 23 June 2015.

[10] Chao Huang, Martin Styner & Hongtu Zhu′ Clustering High-Dimensional Landmark-based Two-dimensional Shape Data, Journal of the American Statistical Association, Apr. 2015.

[11] Hamid Shahriari and Orod Ahmadi, Robust estimation of the mean vector for high-dimensional data set using robust clustering, Econ papers, Journal of Applied Statistics, 2015, vol. 42, issue 6, pages 1183-1205.

[12] Thenmozhi Srinivasan1 and Balasubramanie Palanisamy, Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence, Hindawi, The science world journal, 2015.

[13] N. Toma˘sev, M. Radovanovi´c, D. Mladeni´c, and M. Ivanovi´c, "Hubness-based fuzzy measures for high-dimensional knearest neighbor classification," in Proc. 7th Int. Conf. on Machine Learning and Data Mining (MLDM), 2011, pp. 16–30.

[14] M. Radovanovi´c, A. Nanopoulos, and M. Ivanovi´c, "Timeseries classification in many intrinsic dimensions," in Proc. 10th SIAM Int. Conf. on Data Mining (SDM), 2010, pp. 677–688.

[15] Fionn Murtagh, Pedro Contreras, Random Projection Towards the Baire Metric for High Dimensional Clustering, Springer, Statistical Learning and Data Sciences Lecture Notes in Computer Science Volume 9047, 2015, pp 424-431.
[16] Q. Song, J. Ni, G. Wang, A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data, IEEE Transactions on Knowledge and Data Engineering, 2011.