International Journal of Applied Mathematical Research, *1 (4) (2012) 520-530*
©Science Publishing Corporation
www.sciencepubco.com/index.php/IJAMR

An Approach for Prediction of Diabetic Disease by Using b-Colouring Technique in Clustering Analysis

D. Vijayalakshmi, K. Thilagavathi,

Department of Mathematics, Kongunadu Arts and Science College, Coimbatore - 29

E-mail: viji_kasc@ymail.com, ktmaths@yahoo.com

Abstract

Medical Data mining is the process of extorting hidden patterns from medical data. We propose the work which presents the development of clustering techniques for classifying Pima Indian diabetic database (PIDD). Here, clustering algorithm is used for predicting diabetic disease based on graph b-colouring technique. The proposed technique presents a real representation of clusters by dominant objects that assures the inter cluster disparity in a partitioning and used to evaluate the quality of cluster

Keywords: *b-chromatic, b-colouring, chromatic number, clustering*

1 Introduction

In this work, a colouring method, namely b-colouring, well-adapted for clustering is proposed. The b-colouring is defined as follows: Let $G=(V,E)$ [1] be an undirected connected and simple graph with vertex set V and edge set E. the b-colouring of G is a vertex colouring function c from V to the set of colours $\{1,2,3,\ldots k\}$ such that: For each pair of adjacent vertices $(v_i,v_j) \epsilon E$, $c(v_i) \neq c(v_j)$(proper colouring) In each colour class, there exists at least one vertex having neighbors in all other colours classes. Such a vertex is called a dominating vertex. It allows building a fine partition of the data set where the number of clusters is not set in advance. In fact, the purpose of this approach is to use dissimilarities between all objects to find a partition of the data set in clusters where the cluster separation is achieved simultaneously with the cluster consistency. This approach exhibits two important features it is robust in the presence of outliers and it identifies each cluster by at least one dominant object which guarantees the disparity between clusters. Data analysis motivates many computing applications, either in a planning phase or as part of their on-line operations. Analysis of data is a process of examining, cleaning, transforming, and modeling data with the goal of emphasizing useful information, providing conclusion, and giving correct decision making. The paper is organized as follows: In Section II, Clustering and its importance was discussed. In Section III, diabetes and its classification were discussed. In Section IV, Algorithms for Classifying the Pima Indian diabetic dataset was analyzed. Applications of b-colouring technique in clustering were discussed in Section V. Section VI deals with experimental analysis. Finally the conclusion was given in Section VII.

2 Importance of Clustering

2.1 What is clustering?

Clustering is one of the most important research areas in the field of data mining. Clustering is a data mining technique used to creating group of objects based on their features in such a way that the objects belonging to the same groups are similar and those belonging to the different groups are dissimilar. The main advantage of clustering is that interesting patterns and structures can be found directly from very large data sets without the background knowledge. Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

2.2 Applications of Graph colouring in clustering

Graph colouring techniques are highly utilized in computer science applications, especially in research areas of computer science such as data mining, image segmentation, clustering, image capturing, networking etc., For example a data structure can be designed in the form of tree which in turn utilized vertices and edges. Similarly modeling of network topologies can be done using graph concepts. In the same way the most important concept of graph colouring is utilized in resource allocation, scheduling. Also, paths, walks and circuits in graph theory are used in tremendous applications say traveling salesman problem, database design concepts, resource networking. This leads to the development of new algorithms and new theorems that can be used in tremendous application especially in clustering. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees It is a new framework for representing a set of multi-dimensional gene expression data as a Minimum Spanning Tree (MST), a concept from the graph theory. A key property of this representation is that each cluster of the expression data corresponds to one sub tree of the MST, which rigorously converts a multi-dimensional clustering problem to a tree partitioning problem. Graph theoretic approach to image segmentation. The data to be clustered or segmented by an undirected graph *G* with are capacities assigned to the similarity between the linked vertices. Clustering is achieved by removing the arcs of *G* to form mutually exclusive sub graphs such that the largest inter-sub graph maximum flow is minimized**.** Clustering Technique with an Example Cluster analysis is the collection of a patterns into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. An example of clustering is depicted in Figure 1. The input patterns are shown in Figure 1(a), and the desired clusters are shown in Figure 1(b). Here, points belonging to the same cluster are given the same label. The variety of techniques for representing data, measuring similarity between data elements, and grouping data elements has produced a rich and often confusing assortment of clustering methods.

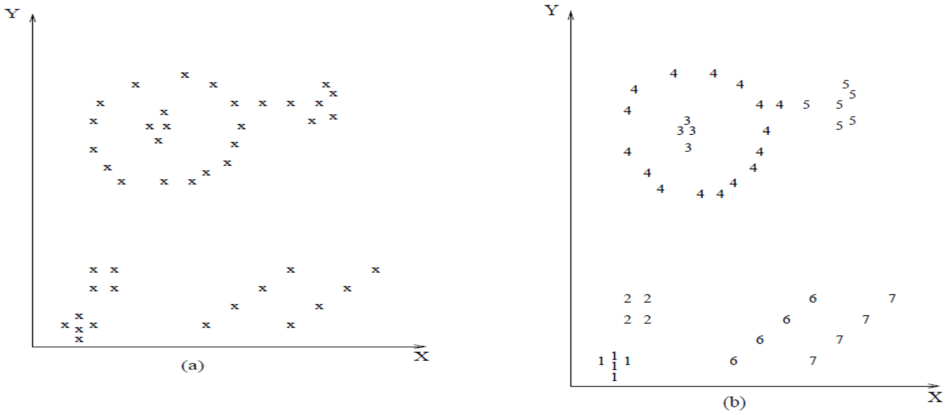

Figure 1: Data clustering

It is important to understand the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification). In supervised classification, we are provided with a collection of labeled (pre classified) patterns; the problem is to label a newly encountered, yet unlabeled, pattern. Typically, the given labeled (training) patterns are used to learn the descriptions of classes which in turn are used to label a new pattern. In the case of clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. In a sense, labels are associated with clusters also, but these category labels are data driven; that is, they are obtained solely from the data. Clustering is useful in several exploratory pattern-analysis, grouping, decision- making, and machine-learning situations; including data mining, document retrieval, image segmentation, and pattern classification. However, in many such problems, there is little prior information (e.g., statistical models) available about the data, and the decision-maker must make as few assumptions about the data as possible. It is under these restrictions that clustering methodology is particularly appropriate for the exploration of interrelationships among the data points to make an assessment (perhaps preliminary) of their structure. Limitations of Clustering algorithms most clustering algorithms are generally based on two popular clustering techniques, namely partitioning and hierarchical clustering [2]. The hierarchical clustering algorithms build a cluster hierarchy or, in other words, a tree of clusters, (dendrogram) whose leaves are the data points and whose internal nodes represent nested clusters of various sizes [3].Hierarchical clustering methods can be further subdivided into agglomerative and divisive. Given *n* objects to be clustered, the agglomerative (bottom-up) methods begin with *n* clusters (i.e., all objects are apart). In each step, the two most similar clusters are identified and merged. This process continues until all objects are clustered into one group. On the other hand, the divisive (top-down) methods start from one cluster containing the *n* objects and split it until *n* clusters are obtained. Once two objects have been merged by an agglomerative algorithm, they will always be in the same cluster; if a divisive method separates two objects, they will never belong to the same cluster. Among these hierarchical methods, some authors have proposed to use graph colouring techniques for the classification purpose. In, the authors propose a divisive classification method based on dissimilarity tables, where the iterative algorithm consists, at each step, in finding a partition by subdividing the class with the largest diameter2 into two classes in order to exhibit a new partition with minimal diameter. The subdivision is obtained by a 2-colouring of the vertices of the maximum spanning tree built from the dissimilarity table. The derived classification structure is a hierarchy. In contrast, given the number k of partitions to be found, partitioning clustering algorithms try to find the best k partitions of the n objects. The partitioning algorithm typically starts with an initial partition of a data set, and then uses an iterative control strategy to optimize an objective function. It is very often the case that the *k* clusters found by a partitioning method are of higher quality (i.e., more similar) than the *k* clusters produced by a hierarchical method [4]. Owing to this property, developing partitioning methods has been one of the main focuses of cluster analysis research. Indeed, many partitioning methods have been developed, some based on k-means [5], some on k-medoid [4], etc. The most popular partitioning algorithm used in scientific and industrial applications is k-means. In k-means, each cluster is represented by its centroid, which is the average of the data vectors in a cluster. The k-means tries to minimize E, which is the average dissimilarity from any object in the data set to the centroid of its cluster. Although k-means often finds the local optimum, it works well when objects within each cluster are quite close to each other while the objects from different clusters are not. It also displays some limitations on attribute types (it does not work well with symbolic attributes) and it is very sensitive to (i) the selection of the initial partition and (ii) to the presence of outliers.

3 Diabetes and its classification

World Health Organization (WHO) report had shown a marked increase in the number of diabetics and this trend is expected to grow in the next couple of decades. In the International Diabetes Federation Conference 2003 held in Paris, India was labeled, as "Diabetes Capital of the World," as of about 190 million diabetics worldwide, more than 33 million is Indians. The worldwide figure is expected to rise to 330 million, 52 million of them Indians by 2025, largely due to population growth, ageing, urbanization, unhealthy eating habits and a sedentary lifestyle. Diabetes mellitus is a disease in which the body is unable to produce or unable to properly use and store glucose (a form of sugar). Glucose backs up in the bloodstream causing one’s blood glucose or "sugar" to rise too high. There are two major types of diabetes. In type 1 (also called juvenile-onset or insulin-dependent) diabetes, the body completely stops producing any insulin, a hormone that enables the body to use glucose found in foods for energy. People with type 1 diabetes must take daily insulin injections to survive. This form of diabetes usually develops in children or young adults, but can occur at any age. Type 2 (also called adult-onset or non-insulin-dependent) diabetes results when the body doesn’t produce enough insulin and/or is unable to use insulin properly (insulin resistance). This form of diabetes usually occurs in people who are over 40, overweight, and have a family history of diabetes, although today it is increasingly occurring in younger people, particularly adolescents. Type II Diabetes (not depending on insulin) is the most common form of diabetes (90 to 95 per cent) and occurs primarily in adults but is now also affecting children and young adults. Type I Diabetes (insulin-dependent) affects predominately children and youth, and is the less common form of diabetes (5 to 10 percent). The major risk factors for diabetes include obesity, high cholesterol, high blood pressure and physical inactivity. The risk of developing diabetes also increases, as people grow older. People who develop diabetes while pregnant (a condition called gestational diabetes) are more likely to develop full-blown diabetes later in life. Poorly managed diabetes can lead to a host of long-term complications among these are heart attacks, strokes, blindness, kidney failure, blood vessel disease [6] [7].

4 Clustering Algorithms for Classifying Pima Indian Diabetic Dataset

A lot of research work has been done on various medical data sets including Pima Indian diabetes dataset. Classification accuracy achieved for Pima Indian diabetes dataset using 22 different classifiers is given in [8] and using 43 different classifiers is given in [9]. The performance of proposed cascaded model (k-means+KNN) is compared with [8] and [9]. The accuracy of most of these classifiers is in the range of 66. 6% to 77.7%. Hybrid K-means and Decision tree [10] achieved the classification accuracy of 92.38% using 10 fold cross validations, cascaded learning system based on Generalized Discriminate analysis (GDA) and Least Square Support Vector Machine (LS_SVM), showed accuracy of 82.05% for diagnosis of Pima dataset [11]. Further authors have achieved classification accuracy of % 72.88 using ANN, 78.21% using DT_ANN where decision tree C4.5 is used to identify relevant features and given as input to ANN [12], 79.50% using Cascaded GA_CFS_ANN, relevant feature identified by Genetic algorithm with Correlation based feature selection is given as input to ANN [13], 77.71% using GA optimized ANN, 84.10% using GA optimized ANN with relevant features identified by decision tree and 84.71% with GA optimized ANN with relevant features identified by GA_CFS[14].

5 Applications of b-colouring Technique in Clustering

5.1 Why Graph b- colouring based clustering?

The graph b colouring based clustering reduced the partitioning problem of a data set into p classes with minimal diameter, to the minimal colouring problem of a superior threshold graph. The edges of this graph are the pairs of vertices distanced from more than a given threshold. In such a graph, each colour corresponds to one class and the number of colours is minimal. Therefore, this method tends to build a partition of the data set with effectively compact clusters but which does not give any importance to the cluster-separation. Graph colouring is used to characterize some properties of graphs. A b-colouring of a graph *G* (using colours *1,2,...,k*) is a colouring of the vertices of *G* such that (i) two neighbors have different colours (proper colouring) and (ii) for each colour class there exists a dominating vertex which is adjacent to all other *k-1* colour classes. In this work, based on a b-colouring of a graph, we propose a clustering technique. Additionally, we provide a cluster validation algorithm. This algorithm aims at finding the optimal number of clusters by evaluating the property of colour dominating vertex.

5.2 How b-colouring is applied in clustering?

The graph b-colouring is an interesting technique that can be applied to various domains. The proper b-colouring problem is the assignment of colours (classes) to the vertices of one graph so that no two adjacent vertices have the same colour, and for each colour class there exists at least one dominating vertex which is adjacent (dissimilar) to all other colour classes. This paper presents a new graph b-colouring framework for clustering heterogeneous objects into groups. A number of cluster validity indices are also reviewed. Such indices can be used for automatically determining the optimal partition. The concept b-colouring is framework for clustering heterogeneous objects into groups. A number of cluster validity indices are also reviewed. Such indices can be used for automatically determining the optimal partition. The proposed approach has interesting properties and gives good results on benchmark data set as well as on real medical data set.

6 Experimental Analysis on Pima Indian Diabetic Dataset

6.1 Pima Indian diabetic dataset

The Pima Indian diabetes database, donated by Vincent Sigillito, is a collection of medical diagnostic reports of 768 examples from a population living near Phoenix, Arizona, USA. The paper dealing with this data base [15] uses an adaptive learning routine that generates and executes digital analogs of perceptron-like devices, called ADAP. They used 576 training instances and obtained a classification of 76% on the remaining 192 instances. The samples consist of examples with 8 attribute values and one of the two possible outcomes, namely whether the patient is tested positive for diabetes (indicated by output one) or not (indicated by two). The database now available in the repository has 512 examples in the training set and 256 examples in the test set. The attribute vectors of these examples are in Table1.

6.2 Clustering using b-colouring technique

The clustering based on the b-colouring of a graph is used for clustering Pima Indian diabetic dataset. Consider the data to be clustered as an undirected edge-weighted graph with no self-loops $G = (V, E)$, where $V = \{v_1,...,v_n\}$ is the vertex set and $E = V \times V$ is the edge set. Vertices in G correspond to objects, edges represent neighborhood relationships, and edge-weights reflect dissimilarity between pairs of linked vertices. The graph G is traditionally represented with the corresponding weighted dissimilarity matrix, which is the $n \times n$ symmetric matrix $D=\{d_{i,j}| \; v_i,v_j \in V\}$. A common informal definition states that “a cluster is a set of entities which are similar, and entities from different clusters are not similar”.

Hence, a cluster should satisfy two fundamental conditions: (1) it should have high internal homogeneity; (2) there should be high heterogeneity between entities within one cluster and those from other clusters.

These two conditions amount to saying that the weights on the edges within a cluster should be small, and those on the edges connecting the cluster nodes to the external ones should be large.

Table 1: Attributes

Attribute	**Type**
Number of times pregnant	continuous
Plasma glucose concentration	continuous
Diastolic blood pressure (mm Hg)	continuous
Triceps skin fold thickness (mm)	continuous
2-Hour serum insulin (mu U/ml)	continuous
Body mass index [weight in kg/(height)]	continuous
Diabetes pedigree function	continuous
Age (years)	continuous
Class variable	Binary

Table 2: A part of Dissimilarity table in Pima Indian diabetes dataset

V_i	1	2	3	4	5	6	7	8	9
1	0								
2	0.20	0							
3	0.10	0.30	0						
4	0.10	0.20	0.25	0					
5	0.20	0.20	0.15	0.40	0				
6	0.20	0.20	0.20	0.25	0.65	0			
7	0.15	0.10	0.15	0.10	0.10	0.75	0		
8	0.10	0.20	0.10	0.10	0.05	0.05	0.05	0	
9	0.40	0.075	0.15	0.15	0.15	0.15	0.15	0.15	0

Table 3: Accuracy when *k=4*

Classifier	**accuracy %**
KNN with all samples , *k=4*	74.4826
K means Cluster all samples, *k=4*	84.76
A graph b-colouring approach all samples, *k=4*	93.676

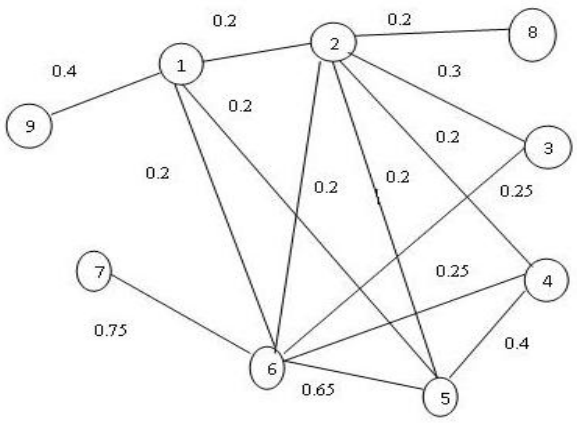

Figure 2: A threshold graph with θ = 0.15 for data in Table 2

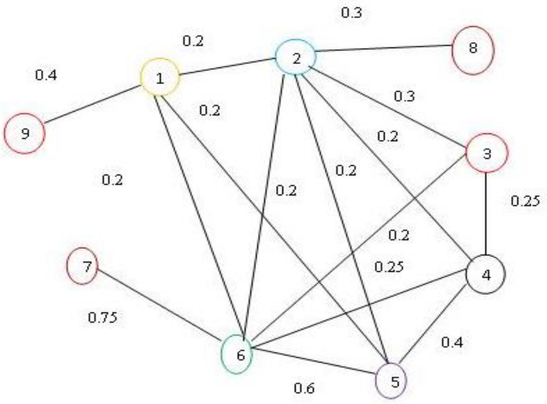

Figure 3: Initializing colours of vertices with maximal colours in Figure 2

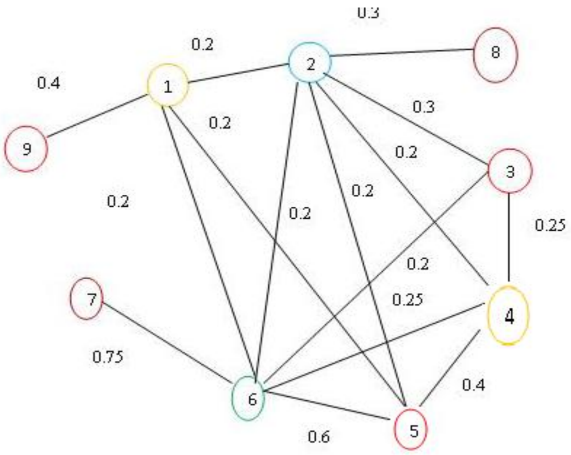

Figure 4: A b colouring graph constructed from Fig 3 by removing colours without any dominating vertex

The dataset is divided into training data and test data using 60-40 ratio. Experiments were carried out for different values of *k* ranging from 1 to 15. Table 3 shows the improvement in accuracy Diabetic data set using proposed method with *k = 4*.

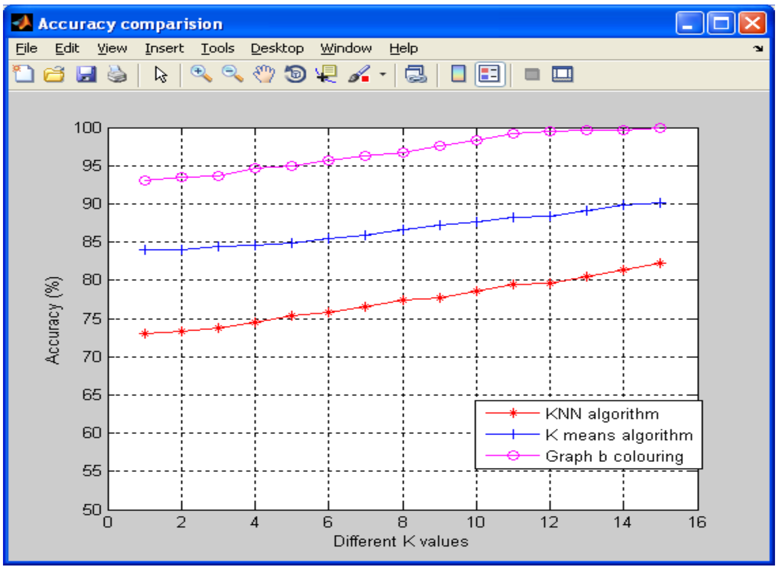

Figure 5: Accuracy Comparison

7 Conclusion

In this work, a clustering algorithm based on a graph b-colouring technique was used to cluster Pima Indian diabetic dataset. We have implemented, performed experiments, and compared our approach with KNN Classification and K-means clustering. The results show that the clustering based on graph colouring outperformance than the other clustering approach in terms of accuracy and purity. The proposed technique offers a real representation of clusters by dominant objects that guarantees the inter cluster disparity in a partitioning and used to evaluate the quality of cluster.

References

[1] B. Effantin , Hamamache Kheddouci, “A Distributed algorithm for b-colouring of a graphs”, lecture Notes in Computer science, Vol 3301, (2006), pp 430-438.

[2] Jain, A.K., M.N. Murty, and P.J. Flynn, “Data Clustering A Review”, ACM Computing Surveys, Vol. 31, (1999), 264-323.

[3] Guha, S., R. Rastogi, and K. Shim, “CURE: An efficient clustering algorithm for large databases”, Proceedings of the ACM SIGMOD Conference, Seattle, WA, (1998), 73-84.

[4] NG, R. and J. Han,”CLARANS: a method for clustering objects for spatial data mining”, IEEE Transactions on Knowledge and Data Engineering, 14(5), (2002), 1003-1016.

[5] Hartigan, J. and M. Wong, “Algorithm AS136: A k-means clustering algorithm”, Journal of Applied Statistics, Vol. 28, (1979), 100-108.

[6] Editorial, “Diagnosis and Classification of Diabetes Mellitus, American Diabetes Association, Diabetes Care”, Vol 27, Supplement 1, (Jan 2004).

[7] The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus, “Follow up report on the Diagnosis of Diabetes Mellitus. Diabetic Care” 26,pp.3160- 3167, (2003).

[8] Michie, D., Spiegelhalter, D. J., & Taylor, C. C., “Machine learning, neural and statistical classification”. 1994.

[9] Humar, K., & Novruz, A, “Design of a hybrid system for the diabetes and heart diseases” Expert Systems with Applications, 2008, 35, 82–89.

[10] B.M Patil, R.C Joshi, Durga Tosniwal, “Hybrid Prediction model for Type-2 Diabetic Patients”, Expert System with Applications, 37, 2010, 8102-8108.

[11] Polat, K., Gunes, S., & Aslan, A., “A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine” Expert Systems with Applications, 2008,34(1), 214–221.

[12] Asha Gowda Karegowda ,MA.Jayaram , “Integrating Decision Tree and ANN for Categorization of Diabetics Data”, International Conference on Computer Aided Engineering, December 13-15, 2007, IIT Madras, Chennai, India.

[13] Asha Gowda Karegowda and M.A. Jayaram, “Cascading GA & CFS for Feature Subset Selection in Medical Data Mining” , International Conference on IEEE International Advance Computing Conference (IACC’09) on March 6-7, 2009, Thapar University, Patiala, Punjab India.

[14] Asha Gowda Karegowda , A.S. Manjunath , M.A. Jayaram, “Application Of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis Of Pima Indians Diabetes”, International Journal on Soft Computing (IJSC), Vol.2, No.2, May 2011.

[15] Smith, J.,W., Everhart, J.,E., Dickson, W.,C., Knowler, W.,C. and Johannes, R.,S., “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in Proceedings of the Symposium on Computer Applications and Medical Care”, IEEE Computer Society Press, 261-265, 1988.