International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249–8958, Volume-1 Issue-3, February 2012

Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients

Asha Gowda Karegowda , M.A. Jayaram, A.S. Manjunath

Abstract*: *Medical Data mining is the process of extracting hidden patterns from medical data. This paper presents the development of a hybrid model for classifying Pima Indian diabetic database (PIDD). The model consists of three stages. In the first stage, K-means clustering is used to identify and eliminate incorrectly classified instances. In the second stage Genetic algorithm (GA) and Correlation based feature selection (CFS) is used in a cascaded fashion for relevant feature extraction, where GA rendered global search of attributes with fitness evaluation effected by CFS. Finally in the third stage a fine tuned classification is done using K-nearest neighbor (KNN) by taking the correctly clustered instance of first stage and with feature subset identified in the second stage as inputs for the KNN. Experimental results signify the cascaded K-means clustering and KNN along with feature subset identified GA_CFS has enhanced classification accuracy of KNN. The proposed model obtained the classification accuracy of 96.68% for diabetic dataset.

Index Terms*: *Genetic algorithm, Correlation based feature selection ,K-nearest neighbor, K-means clustering , Pima Indian Diabetics.

I. INTRODUCTION

The data mining functionalities are used to specify the kind of patterns to be found in the data-mining task. The data mining functionalities mainly include association rule mining, classification, prediction & clustering. Association analysis is used for discovering interesting relations between variables in large databases, w AUTHOR PROFILEhich in given in the form of rules to user. Classification predicts the class labels. Prediction is used to access the value of an attribute that a given sample is likely to have. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

Manuscript published on 28 February 2012.
* Correspondence Author (s)
Asha Gowda Karegowda*, Dept. of Master of Computer Applications, Siddaganga Institute of Technology, Tumkur, India, 9844327268, ashagksit@gamil.com
M.A. Jayaram, Dept. of Master of Computer Applications, Siddaganga Institute of Technology, Tumkur, India, 8095992902, jayaramdps@gmail.com
A.S. Manjunath, Dept. of Computer Science and Engg., Siddaganga Institute of Technology, Tumkur, India, 9845141040, asmanju@gmail.com

Classification is supervised learning algoirthms in contrasts with clustering, which are unsupervised learning algorithm [1]. Classification is a supervised model, which maps or classifies a data item into one of several predefined classes. Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. Typically the learned model is represented in the form of classification rules, decision trees, or mathematical formulae. In the second step the model is used for classification. The most common classification data mining techniques are Case-Based Reasoning, Decision tree, Backpropagation neural network, Radial basis neural network, Bayesian classification, Rough set Approach, Fuzzy Set Approaches, K-nearest Neighbor classifiers.The classifiers are of two types. (a) Instance based or lazy learners in which it store all of training samples and do not build a classifier until a new sample with no class label needs to be classified. K-nearest neighbor(KNN), Case-based reasoning (CBR) are instance-based classifiers. (b) Eager learning methods construct classification model using training data which is tested using test data. Decision tree, Backpropagation neural network, Radial basis neural network using eager learning methods. In this paper a cascaded K-means clustering and k-nearest neighbor classfication algorithm has been used to categorize diabetics patients. Literature survey of classification of diabetic data set is briefed in section II. For the sake of completeness KNN classifier and K-mean clustering have been briefly explained in section III and IV. Feature extraction using GA_CFS and working of cascaded K-means clustering and KNN classifier is explained in section V, followed by results and conclusion in section VI and VII respectively.

II. RELATED WORK ON CLASSIFICATION OF DIABETIC DATA SET

A. Diabetes

A 199

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 are 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-dependant) 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 [2], [3].

B. Literature review of classfication of 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 [4] and using 43 different classifiers is given in [5]. The performance of proposed cascaded model (k-means+KNN) is compared with [4] and [5]. The results of [5] and [4] are shown in Table 1 and Table 2 respectively. The accuracy of most of these classifiers is in the range of 66. 6% to 77.7%. Hybrid K-means and Decision tree [6] 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 [7]. 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 [8], 79.50% using Cascaded GA_CFS_ANN, relevant feature identified by Genetic algorithm with Correlation based feature selection is given as input to ANN [9], 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[10].

III. K-NEAREST NEIGHBOR ALGORITHM

KNN are instance-based or lazy learners[1]. It delays the process of modeling the training data until it is needed to classify the test samples. It can be used both for classification and prediction.

The training samples are described by n-dimensional numeric attributes. The training samples are stored in an n-dimensional space. When a test sample (unknown class label) is given, the k-nearest neighbor classifier searches the k tranining samples which are clsoest to the unknown sample. Closeness is usually defined in terms of Eucliedean distance. The euclidean distance is between two points P(p1,p2, …. Pn) and Q(q1,q2,…. qn) given by equation 1.

$$d(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

k-nearest neighbor classfication algorithm.

1. Let k be number of nearest neighbors and D be the set of training samples(y_j).
2. for each test sample x_i do
   compute d(x_i , y_j) using Euclidean distance for every sample y_j of D
3. Select the k closeset training samples y_j (neighbours) to test sample x_i
4. Classify the sample x_i based on majority class among its nearest neighbors.
5. end for

Some of the advantages of KNN are a)it is very to simple to implement and easy to justify the outcome of KNN. Although KNN has this advantages, it has some disadvantages such as: a) high Computation cost since it needs to compute distance of each test instance to all training samples b) requires large memory proportional to the size of training set c) Low accuracy rate in multidimensional data sets with irrelevant features d) there is no thumb rule to determine value of parameter K (number of nearest neighbors).

IV. K-MEANS CLUSTERING

K-means is one of the simplest unsupervised learning algorithms and follows partitioning method for clustering. K-means algorithm takes the input parameter, k as number of clusters and partitions a dataset of n objects into k clusters, so that the resulting objects of one cluster are dissimilar to that of other cluster and similar to objects of the same cluster. In k-means algorithms begins with randomly selected k objects, representing the k initial cluster center or mean. Next each object is assigned to one the cluster based on the closeness of the object with cluster center. To assign the object to the closest center, a proximity measure namely Euclidean distance is used that quantifies the notion of closest. After all the objects are distributed to k clusters, the new k cluster centers are found by taking the mean of objects of k clusters respectively. The process is repeated till there is no change in k cluster centers. K-means
algorithm aims at minimizing an objective function, namely sum of squared error (SSE).
SSE is defined as

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \qquad (1)$$

Where E is sum of the square error of objects with cluster means for k cluster. p is the object belong to a cluster Ci and mi is the mean of cluster C_i .The time complexity of K-means is O (t*k*n) where t is the number of iterations, k is number of clusters and n is the total number of records in dataset.

K-means partitioning algorithm:
Input is k is the number of clusters, D is input data set
Output is k clusters.

1. Randomly choose k objects from D as the initial cluster centers.
2. Repeat
3. Assign each object from D to one of k clusters to which the object is most similar based on the mean value of the objects in the cluster.
4. Update the cluster means by taking the mean value of the objects for each of k cluster.
5. Until no change in cluster means/ min error E is reached.

V. CATEGORIZATION OF DIABETIC PATIENTSUSING CASCADED PROPOSED MODEL

A. *Data preprocessing.*

The PIMA diabetic database consist of two categories in the data set (i.e Tested positive , Tested Negetive) each having 8 features :Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function and Age (years]. The data is availed from UCI Machine Learning Repository available at http://www.ics.uci.edu/~mlearn/MLRepository.html The data processing techniques, when applied prior to mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining. Data preprocessing is a significant step in the knowledge discovery process, since quality decisions must be based on quality data. A total of 768 cases are available in PIDD. 5 patients had a glucose of 0, 11 patients had a body mass index of 0, 28 others had a diastolic blood pressure of 0, 192 others had skin fold thickness readings of 0, and 140 others had serum insulin levels of 0. After deleting these cases there were 392 cases with no missing values (130 tested positive cases and 262 tested negative) [11].

B. *Working of Proposed method*

In the first stage of proposed model, simple K-means clustering (with k = 2) of Weka tool, is applied to 392 diabetic patients samples as obtained in section V A. The wrongly classified samples are eliminated to get final 299 samples.

The accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features .In the second stage , relevant features are identified by cascaded GA_CFS. In supervised learning, feature selection aims to maximize classification accuracy. Mining on the reduced set of attributes has following benefits a)It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand b)It enhances the classification accuracy and c)It reduces classifier- learning time. GA is a stochastic general search method, capable of effectively exploring large search spaces, which is usually required in case of attribute selection. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Correlation coefficients is used to estimate correlation between subset of attributes and class, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation [12]. Authors have used GA_CFS to improve performance Artificial neural network for categorization of diabetic patients [9] and Bayesian, Naïve bayes, KNN, RBNF, Decision tree classifier for categorization of medical data sets: Breast cancer, Heart, Diabetic and Dermatology datasets [13]. Finally in the third stage, the correctly classified samples from first stage and the relevant features identified in the second stage were given as input to Weka KNN classifier using 70-30 ratio partitioning method (training-test). Classification accuracy and Kappa statistics are used for performance evaluation of proposed classifier. When Kappa statistics K equals 1 , it means that there is complete agreement between the classifier output and the expected output (real world output). Kappa is always less than or equal to 1. A value less than 1 implies less than perfect agreement between the classifier output and the real world expected output.

VI. RESULTS

In this paper k-nearest neighbor (k-NN) classfier has been used for classification of standard medical database namely PIMA diabetic. Incorrect labeled instance are eliminated using K-means clusering followed by feature extraction using GA_CFS. The resultant dataset is divided into training data and test data using 60-40 ratio. Experiments were carried out for different values of k ranging k from 1 to 15. Table 3 shows the improvement in accuracy Diabetic data set using proposed method without feature selection is 95.56% with k = 5. Further with feature selection using GA_CFS, the accuracy of proposed method is found to be 96.67 % with k= 5, sensitivity and specifivity is 100 and 88 respectively. Kappa statistics K value for proposed cascaded model(k-means + classifier with feature slection) is equal to 0.9137 , (which is almost equal to 1) clearly illustrats the match between the proposed cascased classifier and and the real world output .

VII. CONCLUSIONS

The proposed cascaded model with K-means clustering and k-nearest neighbor classifier has been successfully used for markable improved classification of PIMA diabetic dataset. Further results

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Method	Accuracy %	Reference
Proposed Model K-means + KNN,k=5	96.68	This paper
Logdisc	77.7	Statlog
IncNet	77.6	Norbert Jankowski
DIPOL92	77.6	Statlog
Linear Discr. Anal.	77.5–77.2	Statlog; Ster & Dobnikar
SMART	76.8	Statlog
GTO DT (5 · CV)	76.8	Bennet and Blue
kNN, k = 23, Manh, raw, W	76.7 ± 4.0	WD-GM, feature weighting 3CV
kNN, k = 1:25, Manh, raw	76.6 ± 3.4	WD-GM, most cases k = 23
ASI	76.6	Ster & Dobnikar
Fisher discr. analysis	76.5	Ster & Dobnikar
MLP + BP	76.4	Ster & Dobnikar
MLP + BP	75.8 ± 6.2	Zarndt
LVQ	75.8	Ster & Dobnikar
LFC	75.8	Ster & Dobnikar
RBF	75.7	Statlog
NB	75.5–73.8	Ster & Dobnikar; Statlog
kNN, k = 22, Manh	75.5	Karol Grudzin´ ski
MML	75.5 ± 6.3	Zarndt
SNB	75.4	Ster & Dobnikar
BP	75.2	Statlog
SSV DT	75.0 ± 3.6	WD-GM, SSV BS, node 5CV MC
kNN, k = 18, Euclid, raw	74.8 ± 4.8	WD-GM
CART DT	74.7 ± 5.4	Zarndt
CART DT	74.5	Stalog
DB-CART	74.4	Shang & Breiman
ASR	74.3	Ster & Dobnikar
SSV DT	73.7 ± 4.7	WD-GM, SSV BS, node 10CV strat
C4.5 DT	73.0	Stalog
C4.5 DT	72.7 ± 6.6	Zarndt
Bayes	72.2 ± 6.9	Zarndt
C4.5 (5 · CV)	72.0	Bennet and Blue
CART	72.8	Ster & Dobnikar
Kohonen	72.7	Statlog
kNN	71.9	Ster & Dobnikar
ID3	71.7 ± 6.6	Zarndt
IB3	71.7 ± 5.0	Zarndt
IB1	70.4 ± 6.2	Zarndt
kNN, k = 1, Euclides, raw	69.4 ± 4.4	WD-GM
kNN	67.6	Statlog
C4.5 rules	67.0 ± 2.9	Zarndt
OCN2	65.1 ± 1.1	Zarndt
QDA	59.5	Ster, Dobnikar

Table 1. The Classification accuracy proposed model compared with different machine learning methods on Pima Indian diabetes dataset (Humar &Novruz,2008)

showed that the performance of cascaded model can be further improved by selecting appropriate features using GA_CFS. Performance analysis was done for 70-30 ratio (training-test) partitioning method using Classification accuracy and Kappa statistics measures. The classification accuracies obtained by the proposed cascaded K_means clustering and K-NN classifier is one of the best results compared with the results reported in the literature.

Table2. The Classification accuracy of different machine learning methods on Pima Indian diabetes dataset (Michie et al.,1994)

Machine learning method	Accuracy%
Discrim	77.5
Quaddisc	73.8
Logdisc	77.7
SMART	76.8
ALLOC80	69.9
k-NN	67.6
CASTLE	74.2
CART	74.5
IndCART	72.9
NewID	71.1
AC2	72.4
Baytree	72.9
NaiveBay	73.8
CN2	71.1
C4.5	73
Itrule	75.5
Cal5	75
Kohonen	72.2
DIPOL92	77.6
Backprop	75.2
RBF	75.7
LVQ	72.8

Table 3. The Classification accuracy and kappa statistics k value of proposed method on Pima Indian diabetes dataset.

Classifier	accuracy %	Kappa Statistics
KNN with all samples , k=5	74.7826	0.3809
Proposed cascaded model (with all features) K-means+ KNN,k=5	95.56	0.8864
Proposed cascaded model (Feature selection using GA_CFS) K-means+ KNN,k=5	**96.67**	0.9137

REFERENCES

[1] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques*, San Francisco, Morgan Kauffmann Publishers, (2001)

[2] Editorial, *Diagnosis and Classification of Diabetes Mellitus*, American Diabetes Association, Diabetes Care, vol 27, Supplement 1, (Jan 2004).

[3] *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).

[4] Michie, D., Spiegelhalter, D. J., & Taylor, C. C., *Machine learning, neural and statistical classification.* Ellis Horwood, 1994

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

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Joseph L.Breault, Data Mining Diabetic Databases: Are rough Sets a Useful Addition?, http://www.galaxy.gmu.edu/interface/I01/I2001Proceedings/Jbreault

[12] Mark A. Hall ,Correlation-based Feature Selection for Machine Learning, Dept of Computer science, University of Waikato . http://www.cs.waikato.ac.nz/~mhall/thesis.pdf

[13] Asha Gowda Karegowda, M.A.Jayaram A.S .Manjunath, *Feature Subset Selection using Cascaded GA & CFS: A Filter Approach in Supervised Learning*.,International Journal on Computer Applications (IJCA) Volume 23, No 2,pp 2011 June.

AUTHOR PROFILE

Asha Gowda Karegowda received her MCA degree and M.Phil in Computer Science in 1998 and 2008 from Bangalore University and Madurai Kamraj University, India respectively. She is currently pursuing her Ph.D under Visvesvaraya Technological University, Belgaum,India. She is working as Associate Professor in the Dept of Master of Computer Applications, Siddaganga Instituteof Technology, Tumkur, India. Her research interests are soft computing, image analysis and medical data mining. She has published few papers in International conferences and International Journals.

M.A. Jayaram received his M.Tech in Civil ,MCA degree and PhD from Bangalore University , IGNOU University , and Visvesvaraya Technological University, Belgaum, India in the year 1987, 2002 and 2008 respectively. He is working as Director in the Dept of Master of Computer Applications, Siddaganga Institute of Technology,Tumkur, India. His research interests are soft computing, image analysis and medical data mining. He has published few papers in International conferences and International Journals.

A.S. Manjunath received his M.Tech and PhD in Computer Science 1988 and 2003 from Mysore University and Bangalore University, India respectively. He is working as Professor in the Dept of Computer Science and Engineering, Siddaganga Institute of Technology,Tumkur, India. Her research interests are Embedded Systems and solutions, Networking and communications and soft computing. He has published few papers in International conferences and International Journals.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)