Research Journal of Applied Sciences, Engineering and Technology 14(2): 67-72, 2017
DOI:10.19026/rjaset.14.3991
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2017 Maxwell Scientific Organization Corp.
Submitted: June 28, 2016 Accepted: August 9, 2016 Published: February 15, 2017

Research Article

Unsupervised Discretization: An Analysis of Classification Approaches for Clinical Datasets

[1]M. Shanmugapriya, [1]H.Khanna Nehemiah, [1]R.S. Bhuvaneswaran, [2]Kannan Arputharaj and
[1]J. Jabez Christopher
[1]Ramanujan Computing Centre,
[2]Department of Information Science and Technology, Anna University, Chennai-600025, India

Abstract: Discretization is a frequently used data preprocessing technique for enhancing the performance of data mining tasks in knowledge discovery from clinical data. It is used to transform the real-world quantitative data into qualitative data. The aim of this study is to present an experimental analysis of the variation in performance of two trivial unsupervised discretization methods with respect to different classification approaches. Equal width discretization and equal frequency discretization methods are applied for four benchmark clinical datasets obtained from the University of California, Irvine, machine learning repository. Both the methods were applied for transforming quantitative attributes into qualitative attributes with three, five, seven and ten intervals. Six classification approaches were evaluated using four evaluation measures. From the results of this experimental analysis, it can be observed that there is a variation in the performance of classification algorithms. Accuracy of classification varies with respect to the discretization method used and also with respect to the number of intervals of discretization. Moreover it can be inferred that different classification approaches require different discretization methods. No method can be deemed to be 'the best-suitable' for all applications; hence the choice of an appropriate discretization method depends on data distribution, data interpretability, correlation, classification performance and domain of application.

Keywords: Classification, clinical knowledge-mining, equal frequency discretization, equal width discretization, qualitative data, quantitative data

INTRODUCTION

Data mining is one of the emerging research areas in computer science and information technology. It is a process of extracting patterns, useful information or trends, from retrospective, massive and multidimensional data. Some application areas of data mining techniques for knowledge extraction include business, academics and medicine. Generally, clinical decisions on medical data are often made based on doctor's perception and experience rather than on the knowledge hidden in the database. This might lead to bias, errors and excessive medical costs which affects the quality of service provided to patients. Therefore, Knowledge Discovery in Databases (KDD) is commonly used to improve the quality of service. Integration of KDD process with medical data could reduce medical errors, provide clinical decision support and improve the diagnostic process. Data mining is an important step in KDD and is used for various aspects in the medical domain such as diagnosis, prognosis and decision support (Christopher *et al*., 2015; Jane *et al*., 2016; Nahato *et al*., 2015; Susmi *et al*., 2015; Sweetlin *et al*., 2016). KDD involves the process of finding and interpreting knowledge from data which is described by the following steps: 1)understanding of domain 2) data set selection, 3) data cleaning and preprocessing, 4) data reduction and projection, 5) matching the objective into a data mining method (association rule mining, classification, clustering, regression etc.,), 6) choice of the algorithm for pattern searching, 7) searching for pattern of interest (data mining),8) data interpretation and 9) use of the discovered knowledge (Fayyad *et al*., 1996). Most prior work on KDD focuses on step 7, the data mining step. Data mining applications often involve quantitative data. However many learning algorithms are intended to handle qualitative data (Kohavi and Sahami, 1996). Algorithms that directly deal with quantitative data, learning is less efficient and less effective (Richeldi and Rossotto, 1995). In many

Corresponding Author: H. Khanna Nehemiah, Ramanujan Computing Centre, Anna University, Chennai-600025, India machine learning techniques we need to transform such quantitative data into qualitative data. This process is called data discretization. Data discretization refers to partitioning the data into discrete set of intervals. Each interval is treated as a category.

Data discretization simplifies the original data and also improves the efficiency of prediction. It has several advantages in machine learning and data mining tasks. In particular, it increases the understandability of the classification models that uses rule sets (Liu *et al*., 2002; Fu, 2011). It also reduces the computation time needed for processing the continuous data by dividing data into reduced set of intervals (Mittal and Cheong, 2002). Maslove *et al*. (2013) have evaluated six discretization methods: two supervised methods (minimum descriptive length-based and ChiMerge), three unsupervised methods (equal width, equal frequency and K-means) and one method specific to clinical data with both supervised and unsupervised components (reference range based). They have examined the impact of discretization on three evaluation parameters: accuracy, consistency and simplicity. To evaluate the six discretization methods for accuracy, each of the discretization methods are examined with decision tree and naïve-bayes classification approach. They have evaluated the discretization methods for consistency by deriving the inconsistency count for each discretization experiment. For evaluating simplicity, they count the number of nodes in each decision tree generated by each of the discretization methods. For the evaluation of discretization methods, they use both laboratory data and physiologic data derived from adult patients in the intensive care unit. From the result, they observed that supervised methods were more accurate than unsupervised. Among the supervised methods, equal frequency and K-means performed well.

Yang and Webb (2009) have proved that discretization is an effective technique for probability-based learning. In their study it was inferred that, the effectiveness of discretization in naïve-bayes learning has impact on the performance of naïve-bayes classifiers. They make use of classification error as a performance measure for naïve-bayes classifier. In order to minimize the classification error, they analyze two factors with respect to discretization: 1) Decision boundaries and 2) the error tolerance of probability estimation for each quantitative attribute. From the analysis they conclude that discretization with these factors can affect the classification bias and variance of the classifiers. The effects are named as discretization bias and discretization variance. To manage the discretization bias and variance, they use the concepts called interval frequency and interval number. Moreover, they propose two efficient unsupervised discretization methods called proportional discretization and fixed frequency discretization for managing discretization bias and variance. They evaluate these two methods against four discretization methods for naïve-bayes classifier on 29 benchmark datasets from UCI machine learning repository. The results have demonstrated that the new proposed discretization methods reduce naïve-bayes classification error when compared to current established discretization methods.

This study focuses on two unsupervised discretization techniques: Equal width Discretization and Equal Frequency Discretization. Continuous-valued attributes are discretized into several intervals and the classification performances of five classification approaches are analyzed. The novel observations and findings of the experimental analysis can serve as guiding principles for preprocessing of clinical data.

MATERIALS AND METHODS

The clinical datasets used in this experimental study were selected from the University of California Irvine (UCI) Machine Learning repository. Datasets which contain categorical, discrete and continuous data were chosen. The list of datasets is presented in Table 1. The description about the Cleveland Heart Disease (CHD) dataset, Chronic Kidney Disease (CKD) dataset, Pima Indians Diabetes (PID) dataset and BUPA Liver Disorder (BLD) dataset are presented in Table 2 to 5 respectively. In particular, the PID dataset consists the details of 768 Pima Indian Women.

The continuous-valued attributes in these datasets were discretized using Equal width discretization and equal frequency discretization methods. The former method divides the continuous-valued feature 'f' into k intervals of equal width, where k *is a* user-defined parameter. Thus each interval has a width (w), where $w = (max\text{-}min) / k$ and interval boundaries are $min+w$, $min+2w, \ldots, min+(k\text{-}1)w$. The latter method divides the range of continuous-valued feature into k equally sized bins. Each interval contains approximately same number of instances, where k is a user-defined

Table 1: Datasets used

Dataset	Number of instances	Number of features
Pima Indians Diabetes (PID)	768	9
BUPA Liver Disorder (BLD)	345	7
Cleveland Heart Disease (CHD)	303	76
Chronic Kidney Disease (CKD)	400	25

Table 2: Description ofcleveland heart disease dataset

Attribute name	Description	Type	Range
Age	Age of the person	Discrete	29-77
Sex	Sex of the person	Categorical	0-1
Cp	Chest pain type	Categorical	1-4
Trestbps	Resting blood pressure	Continuous	94-200
Chol	Serum cholestoral	Continuous	126-564
Fbs	Fasting blood sugar	Categorical	0-1
Restecg	Resting electrocardiographic results	Categorical	0-2
Thalach	Maximum heart rate achieved	Continuous	71-202
Exang	Exercise induced angina	Categorical	0-1
Oldpeak	ST depression induced by exercise relative to rest	Continuous	0-6.2
Slope	The slope of the peak exercise ST segment	Categorical	1-3
Ca	Number of major vessels (0-3) colored by flourosopy	Categorical	0-3
Thal	Defect types	Categorical	3-7
Class	Presence /Absence of heart disease	Categorical	0-4

Table 3: Description of chronic kidney disease dataset

Attribute name	Description	Type	Range
Age	Age in years	Discrete	12-90
Bp	Blood pressure	Continuous	50-180
Sg	Specific gravity	Categorical	1.005, 1.010, 1.015, 1.020, 1.025
Al	Albumin	Categorical	0-5
Su	Sugar	Categorical	0-5
Rbc	Red blood cells	Categorical	Normal, abnormal
Pc	Pus cell	Categorical	Normal, abnormal
Pcc	Pus cell clumps	Categorical	Present, not present
Ba	Bacteria	Categorical	Present, not present
Bgr	Blood glucose random	Continuous	22-490
Bu	Blood urea	Continuous	1.5-391
Sc	Serum creatinine	Continuous	0.4-76
Sod	Sodium	Continuous	4.5-163
Pot	Potassium	Continuous	2.5-47
Hemo	Hemoglobin	Continuous	3.1-17.8
Pcv	Packed cell volume	Continuous	9-54
Wc	White blood cell count	Continuous	2200-26400
Rc	Red blood cell count	Continuous	2.1-8
Htn	Hypertension	Categorical	Yes, no
Dm	Diabetes mellitus	Categorical	Yes, no
Cad	Coronary artery disease	Categorical	Yes, no
Appet	Appetite	Categorical	Good, poor
Pe	Pedal edema	Categorical	Yes, no
Ane	Anemia	Categorical	Yes, no
Class	Presence/Absence of kidney disease	Categorical	ckd, notckd

Table 4: Description of pima Indian diabetes dataset

Attribute name	Description	Type	Range
Preg	Number of times pregnant	Discrete	0-17
Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test	Continuous	0-199
Bp	Diastolic blood pressure	Continuous	0-122
Skin	Triceps skin fold thickness	Continuous	0-99
Insulin	2-Hour serum insulin	Continuous	0-846
BMI	Body mass index	Continuous	0-67.1
Pedi	Diabetes pedigree function	Continuous	0-2.42
Age	Age of the person	Discrete	21-81
Class	Diabetes/Non-Diabetes	Categorical	0-1

Table 5: Description of liver disorder dataset

Attribute name	Description	Type	Range
Mcv	Mean corpuscular volume	Continuous	65-103
Alkphos	Alkaline phosphotase	Continuous	23-138
Sgpt	Alamine aminotransferase	Continuous	4-155
Sgot	Aspartate aminotransferase	Continuous	5-82
Gammagt	Gamma-glutamyltranspeptidase	Continuous	5-297
Drinks	Number of half-pint equivalents of alcoholic beverages drunk per day	Continuous	0-20
Class	Diagnosis of disease	Categorical	Present/Absent

parameter. Thus each interval contain *n/k* values, where '*n*' is the total number of instances (records) in the dataset. The discretized data is split into training and testing data. The former is used for obtaining the classifier using an induction algorithm and the latter is used for evaluating the performance of the classifier using performance evaluation measures.

Cross-Validation (CV) with '*k*' folds is a technique whereby the dataset '*D*', is randomly split into *k* folds of approximately equal size. The classifier (model) is trained and tested *k* times. Each time (*k*-1) folds are used for training and the remaining one fold is used for testing. In classification, *k*-fold cross-validation is the best method to use for validating and selecting a classifier (Kohavi, 1995). Associative classifier (CBA), Decision tree classifier (C4.5), Support Vector Machine (SVM), Multi-Layer Perceptron classifier (MLP), Naïve Bayes classifier (NB) and k-Nearest Neighbour classifier (kNN) are validated (Han and Kamber, 2006).

In this experimental study, six trivial classification approaches were used. Each approach differs from the other in two aspects: first, the induction (learning) algorithm used for training the classifier; and second, the knowledge-representation form used to represent the classification model. The six classification approaches are as follows: first, a decision tree classifier (Quinlan, 1986), induced (trained) using the C4.5 algorithm is used. The classifier (knowledge model) is represented in the form of a tree; second, the naïve Bayes classifier uses a probabilistic induction approach and the knowledge model is represented in the form of probabilistic values; third, the Class-Based Associative (CBA) (Liu *et al*., 1998) classifier uses an Apriori-based (Agrawal and Srikant, 1994) classification rule induction approach and the knowledge model is represented in the form of IF-THEN associative classification rules; fourth, the Multilayer Perceptron (MLP) (Rosenblatt, 1958) is induced using a gradient descent-based backpropagation algorithm and the knowledge is represented by a trained feed-forward Neural Network; fifth, the Support Vector Machine (Boser *et al*., 1992) is induced using the Sequential Minimal Optimization (SMO) algorithm and the knowledge model is represented in the form of support vectors and the separating hyper planes; sixth, the K-NN classifier trained using distance-based approach and the classifier is represented in terms of distance measures from neighboring instances. The choice of a classification approach and an appropriate classifier depends on the need and purpose of the classifier for that domain of application. Moreover, factors such as data distribution, entropy of discretization may also be considered.

In this experimental study, four performance evaluation measures were used. The four measures namely, Sensitivity, Specificity, Fmeasure and Accuracy differ in their evaluation focus. Sensitivity is used to evaluate the effectiveness of a classifier to identify positive labels whereas Specificity evaluates how effectively a classifier identifies negative labels. Fmeasurerelates between data's positive labels and those given by a classifier based on per-class average and finally Accuracy evaluates the overall classification efficiency of the classifier.

RESULTS AND DISCUSSION

The evaluation of classification performance of six classification approaches for equal width discretization and equal frequency discretization is presented in Table 6. A discussion on the observations, findings and important inferences are presented below.

For the PID dataset, bayes classifier achieves the highest accuracy of 76.307% for EW discretization with 7 intervals whereas the bayes classifier with 7 intervals for EF discretization yields 73.96%. The highest accuracy for EF discretization for the PID dataset is achieved by C4.5 algorithm (74.867%). Though entropy of the partitions (intervals) are proportional to the number of partitions, a drop in classification accuracy for increase in the number of partitions can be inferred. This accuracy-drop is due to the inter-correlation between the attribute-subset and also the correlation between the attribute and the class attribute. A diminish in the former and a rise in the latter is preferred.

A change in the choice of the attribute selection order or the attribute-subset, for the construction of a decision tree, may result in a variation in classification performance. For example, the highest classification accuracy for EF discretization, for the BLD dataset was achieved by the C4.5 classifier trained using 3 intervals. Moreover, the increase in the number of intervals enhanced the information gain of the individual attributes. But during tree construction, the attribute-subsets for lower levels of the trees yields different combination of attributes; different combination of attributes in the attribute-subsets differ in the level of inter-correlation. Hence a fall in accuracy for EF 10-interval can be observed.

In some scenarios, as the number of intervals increase the number of pure partitions also increase; a pure partition has low entropy and hence it is a desirable characteristic for classification. For example, in the case of the CKD dataset, a drop in accuracy for the five-interval data can be observed. This is due to the disproportionate change in the number of pure partitions for a linear increase in the number of intervals.

Table 6: Classification performance evaluation for Equal Width (EW) and Equal Frequency (EF) discretization methods

			SVM				KNN				C4.5			
Dataset	Method	No. of Intervals	*Acc	Sen	Spec	Fmes	Acc	Sen	Spec	Fmes	Acc	Sen	Spec	Fmes
PID	EW	3	73.823	0.589	0.818	0.606	72.650	0.604	0.792	0.605	73.043	0.563	0.820	0.587
		5	68.628	0.246	0.922	0.353	71.476	0.537	0.810	0.563	73.963	0.529	0.852	0.583
		7	69.137	0.320	0.890	0.412	68.749	0.485	0.796	0.517	74.475	0.607	0.818	0.619
		10	64.985	0.272	0.852	0.350	67.984	0.463	0.796	0.501	73.307	0.428	0.896	0.521
	EF	3	73.438	0.544	0.836	0.583	68.630	0.623	0.720	0.580	75.133	0.566	0.850	0.611
		5	70.965	0.366	0.894	0.462	72.920	0.672	0.760	0.633	74.867	0.570	0.844	0.606
		7	66.541	0.250	0.888	0.338	69.405	0.571	0.760	0.563	73.706	0.489	0.870	0.560
		10	63.156	0.026	0.956	0.047	66.806	0.560	0.726	0.539	72.262	0.462	0.862	0.531
CHD	EW	3	81.151	0.807	0.818	0.821	77.892	0.861	0.679	0.810	76.237	0.808	0.708	0.787
		5	81.473	0.868	0.752	0.837	80.441	0.885	0.709	0.834	77.538	0.827	0.716	0.799
		7	82.817	0.855	0.795	0.847	79.806	0.903	0.673	0.831	77.194	0.819	0.715	0.795
		10	80.204	0.813	0.788	0.819	78.161	0.903	0.636	0.822	79.183	0.850	0.724	0.815
	EF	3	82.452	0.867	0.773	0.845	80.161	0.892	0.694	0.833	77.183	0.826	0.709	0.795
		5	79.839	0.849	0.737	0.822	80.817	0.879	0.723	0.834	77.860	0.844	0.701	0.807
		7	79.484	0.860	0.714	0.823	77.161	0.848	0.680	0.802	77.527	0.838	0.701	0.801
		10	75.505	0.842	0.649	0.791	82.430	0.909	0.722	0.853	78.204	0.839	0.716	0.805
BLD	EW	3	58.277	0.049	0.970	0.085	59.479	0.423	0.720	0.460	55.941	0.177	0.835	0.220
		5	57.689	0.111	0.915	0.157	53.286	0.407	0.625	0.418	55.908	0.166	0.845	0.200
		7	60.017	0.350	0.780	0.419	55.924	0.530	0.580	0.496	54.815	0.386	0.665	0.416
		10	64.916	0.413	0.820	0.495	62.622	0.648	0.610	0.597	61.151	0.438	0.735	0.478
	EF	3	64.655	0.386	0.835	0.473	64.353	0.655	0.635	0.607	70.983	0.522	0.845	0.9
		5	64.429	0.459	0.780	0.512	60.555	0.594	0.615	0.555	66.420	0.407	0.850	0501
		7	60.908	0.366	0.785	0.434	59.975	0.600	0.600	0.554	65.513	0.477	0.785	0.536
		10	62.655	0.271	0.885	0.368	56.496	0.551	0.575	0.517	64.681	0.450	0.790	0.495
CKD	EW	3	96.000	0.936	1.000	0.966	94.750	0.916	1.000	0.955	98.000	0.988	0.967	0.984
		5	92.250	0.896	0.967	0.935	88.750	0.820	1.000	0.898	98.000	0.972	0.993	0.983
		7	94.250	0.948	0.933	0.954	88.750	0.824	0.993	0.897	98.000	0.976	0.987	0.984
		10	94.500	0.948	0.940	0.955	88.750	0.824	0.993	0.898	96.250	0.956	0.973	0.970
	EF	3	95.500	0.944	0.973	0.963	92.250	0.876	1.000	0.931	97.000	0.956	0.993	0.975
		5	96.250	0.960	0.967	0.969	91.500	0.868	0.993	0.926	98.250	0.976	0.993	0.986
		7	95.250	0.956	0.947	0.962	92.000	0.876	0.993	0.930	97.500	0.964	0.993	0.979
		10	92.550	0.936	0.907	0.939	91.750	0.868	1.000	0.928	97.250	0.960	0.993	0.977

			CBA				Bayes				MLP			
Dataset	Method	No. of Intervals	Acc	Sen	Spec	Fmes	Acc	Sen	Spec	Fmes	Acc	Sen	Spec	Fmes
PID	EW	3	65.106	0.000	1.000	0.000	73.561	0.562	0.828	0.591	71.625	0.534	0.814	0.561
		5	63.93	0.422	0.756	0.307	75.267	0.604	0.832	0.625	70.444	0.515	0.806	0.544
		7	65.106	0.000	1.000	0.000	76.307	0.653	0.822	0.656	68.628	0.503	0.784	0.525
		10	67.196	0.759	0.624	0.580	75.533	0.645	0.814	0.646	73.706	0.605	0.808	0.615
	EF	3	67.051	0.873	0.562	0.650	74.880	0.672	0.790	0.649	69.667	0.578	0.760	0.570
		5	63.937	0.248	0.848	0.179	74.228	0.675	0.778	0.645	72.667	0.579	0.806	0.597
		7	65.106	0.000	1.000	0.000	73.968	0.682	0.770	0.646	70.960	0.582	0.778	0.578
		10	65.106	0.000	1.000	0.000	75.005	0.686	0.784	0.654	71.885	0.575	0.796	0.585
CHD	EW	3	74.570	0.673	0.832	0.740	82.796	0.837	0.825	0.844	78.215	0.820	0.737	0.805
		5	74.237	0.624	0.884	0.721	83.462	0.856	0.818	0.853	82.473	0.836	0.810	0.839
		7	70.989	0.577	0.869	0.661	83.462	0.861	0.803	0.852	79.871	0.814	0.781	0.814
		10	68.419	0.518	0.884	0.614	84.140	0.868	0.810	0.857	83.118	0.849	0.809	0.847
	EF	3	76.849	0.739	0.803	0.776	85.108	0.873	0.825	0.867	79.161	0.813	0.765	0.812
		5	67.753	0.542	0.836	0.617	85.430	0.873	0.832	0.868	80.505	0.812	0.796	0.818
		7	67.753	0.542	0.836	0.617	84.118	0.855	0.825	0.857	80.172	0.825	0.774	0.820
		10	67.753	0.542	0.836	0.617	83.129	0.843	0.817	0.846	80.828	0.831	0.780	0.826
BLD	EW	3	57.983	0.000	1.000	0.000	61.193	0.388	0.775	0.451	55.975	0.378	0.695	0.390
		5	57.983	0.000	1.000	0.000	55.050	0.301	0.730	0.348	57.034	0.406	0.690	0.441
		7	57.983	0.000	1.000	0.000	63.479	0.489	0.740	0.517	55.109	0.421	0.645	0.438
		10	57.983	0.000	1.000	0.000	65.269	0.504	0.760	0.547	64.353	0.566	0.700	0.568
	EF	3	57.983	0.000	1.000	0.000	68.706	0.469	0.845	0.560	64.950	0.609	0.680	0.91
		5	57.983	0.000	1.000	0.000	64.092	0.456	0.775	0.514	63.521	0.581	0.675	0.570
		7	57.983	0.000	1.000	0.000	66.109	0.498	0.780	0.549	67.571	0.628	0.710	0.617
		10	57.983	0.000	1.000	0.000	62.353	0.476	0.730	0.513	66.731	0.609	0.710	0.603
CKD	EW	3	97.500	0.972	0.980	0.980	98.000	0.968	1.000	0.983	98.250	0.976	0.993	0.986
		5	96.250	0.976	0.940	0.970	97.250	0.956	1.000	0.977	98.000	0.980	0.980	0.984
		7	96.750	0.976	0.953	0.974	97.250	0.956	1.000	0.977	98.250	0.976	0.993	0.986
		10	96.750	0.976	0.953	0.974	97.750	0.964	1.000	0.981	98.500	0.984	0.987	0.988
	EF	3	96.750	0.976	0.953	0.974	97.000	0.952	1.000	0.975	98.750	0.992	0.980	0.990
		5	96.750	0.976	0.953	0.974	96.250	0.940	1.000	0.968	99.250	0.988	1.000	0.994
		7	96.750	0.976	0.953	0.974	97.750	0.964	1.000	0.981	98.750	0.992	0.980	0.990
		10	96.750	0.976	0.953	0.974	97.750	0.964	1.000	0.981	98.750	0.988	0.987	0.990

* Acc- Accuracy; Sen- Sensitivity; Spec- Specificity; Fmes- Fmeasure/FScore

CONCLUSION

Clinical data usually consist of sensor readings from medical equipments, temperature readings fromthermometers,height and weight measurements from appropriate devices; however representation of such values in an easy human-interpretable form requires the data to be discretized. Improper use of discretization approaches can penalize the efficiency of the data mining tasks such as classification. Moreover appropriate use of discretization, improves the data representation and data interpretability. The observations and findings of this study enable engineers to choose a fitting discretization approach while designing clinical knowledge-based systems. This study is focused on the use of unsupervised approaches for clinical datasets. This study may further be extended by analyzing the effect of many more discretization approaches over various domains. Experimental analysis of more datasets and approaches may yield novel findings which may improve the performance of the systems that use typical data mining tasks.

CONFLICT OF INTEREST

The authors state that there are no financial/relevant interests that influence the development of the manuscript.

REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. Proceeding of the 20th International Conference on Very Large Databases. Santiago, Chile, pp: 487-499.

Boser, B.E., I.M. Guyon and V.N. Vapnik, 1992. A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory, pp: 144-152.

Christopher, J.J., H.K. Nehemiah and A. Kannan, 2015. A clinical decision support system for diagnosis of allergic rhinitis based on intradermal skin tests. Comput. Biol. Med., 65: 76-84.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996. Knowledge discovery and data mining: Towards a unifying framework. Proceeding of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), pp: 82-88.

Fu, T.C., 2011. A review on time series data mining. Eng. Appl. Artif. Intel., 24(1): 164-181.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann, San Francisco, CA, USA.

Jane, N.Y., K.H. Nehemiah and K. Arputharaj, 2016. A temporal mining framework for classifying un-evenly spaced clinical data: An approach for building effective clinical decision-making system. Appl. Clin. Inform., 7(1): 1-21.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceeding of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95), 14: 1137-1143.

Kohavi, R. and M. Sahami, 1996. Error-based and entropy-based discretization of continuous features. Proceeding of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), pp: 114-119.

Liu, B., W. Hsu and Y. Ma, 1998. Integrating classification and association rule mining. Proceeding of the 4th International Conference on Knowledge Discovery and Data Mining, pp: 80-86.

Liu, H., F. Hussain, C.L. Tan and M. Dash, 2002. Discretization: An enabling technique. Data Min. Knowl. Disc., 6(4): 393-423.

Maslove, D.M., T. Podchiyska and H.J. Lowe, 2013. Discretization of continuous features in clinical datasets. J. Am. Med. Inform. Assn., 20(3): 544-553.

Mittal, A. and L.F. Cheong, 2002. Employing discrete bayes error rate for discretization and feature selection tasks. Proceeding of the IEEE International Conference on Data Mining (ICDM-2002), pp: 298-305.

Nahato, K.B., K.N. Harichandran and K. Arputharaj, 2015. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Comput. Math. Method. M., 2015: 1-13.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn., 1(1): 81-106.

Richeldi, M. and M. Rossotto, 1995. Class-driven Statistical Discretization of Continuous Attributes (Extended Abstract). In: Lavrac, N. and S. Wrobel (Eds.), Machine Learning: ECML-95. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 912: 335-338.

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev., 65(6): 386-408.

Susmi, S.J., H.K. Nehemiah, A. Kannan and J.J. Christopher, 2015. A hybrid classifier for leukemia gene expression data. Res. J. Appl. Sci. Eng. Technol., 10(2): 197-205.

Sweetlin, J.D., H.K. Nehemiah and A. Kannan, 2016. Patient-specific model based segmentation of lung computed tomographic images. J. Inform. Sci. Eng., 32(5): 1373-1394.

Yang, Y. and G.I. Webb, 2009. Discretization for naive-bayes learning: Managing discretization bias and variance. Mach. Learn., 74(1): 39-74.