Improve the Classifier Accuracy for Continuous Attributes in Biomedical Datasets using a New Discretization Method

G.Madhu Prof.T.V.Rajinikanth , Prof. A.Govardhan
E-mail: madhu_g@vnrvjiet.in

Motivation

- Many real-world datasets are predominately consist of continuous attributes also called quantitative attributes.
- These type of datasets are unsuitable for certain data mining algorithms that deals only nominal attributes.
- Some classification algorithms such as CLIP and CN2, ID3 are inherently incapable of handling continuous attributes.
- To use such algorithms we need to transform continuous attributes into nominal attributes this process known as 'Discretization'.
- Even though some traditional methods have disadvantages like unbalanced intervals, presence of outliers, also unsupervised, so it ignore the class information.

Proposed Method

- The proposed discretization algorithm is a combination of the concepts Fayyad and Irani discretization algorithms and greedy approach .
- Let sample S= {x_1, x_2, ……., x_n} be the set of real-valued attributes or continuous attributes. Now to discretize the number of continuous attributes in the given dataset, first we need to apply a standardized statistical technique z-score (given below) on dataset. The z-score is defined as follows :

$$z - score\ (S) = \frac{(x_k - \overline{x})}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{k}(x_i - \overline{x})}} \quad \text{.....(1)}$$

- After applying z-score on dataset, we find the minimum and maximum values from dataset. We assume that the minimum value of z-score is 'a' and maximum value of z-score is 'b' from the given dataset.

$$\therefore \quad a \leq x_i < b \quad for\ i = 1, 2, 3, \ldots . n \quad \text{.....(2)}$$

- In order to partition the continuous attributes into a finite number of intervals with all possible value of random variables X.

$$X = [a, b) = \{x / \ a \leq x_i < b \} \quad \text{.....(3)}$$

- After that partition the interval $X = [a, b)$ into a k-equal width bins as follows:

$$[a, b) = \cup_{i=1}^{k-1} B_i = B_0 \cup B_1 \cup B_2 \ldots\ldots\ldots\ldots \cup B_{k-1} \text{.....(4)}$$

- where $\delta = \frac{b-a}{k}$ ..... (5) this represents a width of the each interval in $X = [a, b)$.

- Therefore, the bins are given below:

$$B_0 = [a, a + \delta),$$
$$B_1 = [a + \delta, a + 2\delta) \ldots B_{k-1} = [a + (k-1)\delta, a + k\delta) \text{ ..... (6)}$$

Moreover, empty bins are not allowed in this process.

Algorithm: ZDisc-Discretization

Input: Dataset 'S' consisting of number of rows and column observations, with continuous attributes in the set 'S'.

Output: Discredited dataset, accuracy of the dataset S.

Step 1: Select all the records with continuous values in the data set S, not those attributes in the decision attributes column (i.e. $\subseteq S$).

Step 2: Identify the continuous record R from the set A and apply the normalization technique that is the z-score measure on the dataset S with proposed new discretization method (see in section 3.1).

Step 3: After discretization Split the dataset S into training (Tr) and testing (Ts) sets using a stratified a k- fold cross validation procedure.

Step 4: In Step-3, for each 'k' computes the following procedure:

(i) Build the Classifier (C4.5) using the records obtained from Tr.

(ii) Compute the predicted probabilities (scores) from the C4.5 built in Step (4)-(i) using the test data set Ts.

(iii) Identify and collect the original features from test data set Ts.

Step 5: Repeat the Steps (4)-(i) to Step (4)-(iii) for each fold.

Step 6: Compute the classifier accuracy of the dataset S.

Step 7: RETURN Step (6)

Step 8: STOP

EXPERIMENTS AND RESULTS

Name	#Attributes (R/I/N)	#Examples	#Classes	# Continuous Attributes
Appendicitis (APD)	7(7/0/0)	106	2	07
Cleveland (CLE)	13(13/0/0)	303	5	13
Hepatitis (HEP)	15 (3/3/9)	214	2	10
Pima (PEM)	8(8/0/0)	768	2	08
Breast CancerWis (BCW)	30(30/0/0)	569	2	30

DATASETS USED IN OUR EXPERIMENTS

Table.2. Test classifiers of our algorithm with other discretization methods on Appendicitis

Dataset		C4.5 classifier	SVM classifier
	Discretization Algorithms	10x cross-fold Validation (%Accuracy)	10x cross-fold Validation(%Accuracy)
Appendicitis	ZDISC	**84.90**	**87.73**
	Ameva	83.18	86.09
	Bayesian	86.00	89.63
	CACC	83.18	85.18
	CADD	80.18	80.18
	CAIM	84.09	84.18
	Chi2	85.08	84.00
	Chi-merge	84.09	85.90
	ExtChi2	80.18	80.18
	Fayyad & Irani	83.18	85.09
	PKID	80.18	80.18

Table.3. Test classifiers of our algorithm with other discretization methods on Cleveland

Dataset		C4.5 classifier	SVM classifier
	Discretization Algorithms	10x cross-fold Validation(%Accuracy)	10x cross-fold Validation(%Accuracy)
Cleveland	ZDISC	**57.09**	**57.90**
	Ameva	51.75	56.72
	Bayesian	52.50	56.08
	CACC	50.80	56.70
	CADD	55.11	55.10
	CAIM	53.10	59.05
	Chi2	54.10	58.74
	Chi-merge	54.44	59.07
	ExtChi2	54.75	56.05
	Fayyad & Irani	57.97	57.74
	PKID	56.23	53.86

Table.4. Test classifiers of our algorithm with other discretization methods on Hepatitis

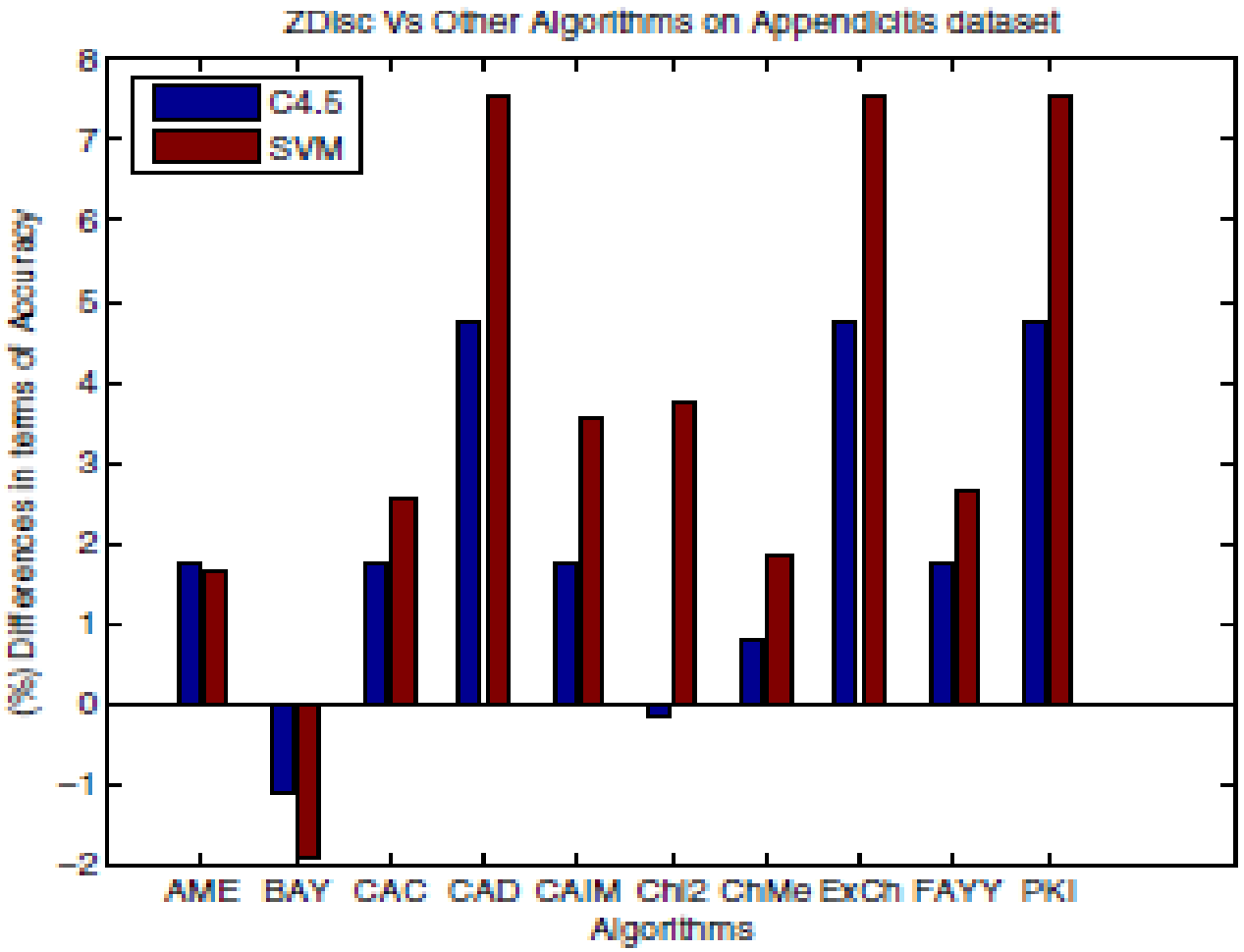
Dataset		C4.5 classifier	SVM classifier
	Discretization Algorithms	10x cross-fold Validation(%Accuracy)	10x cross-fold Validation(%Accuracy)
Hepatitis	ZDISC	**89.95**	**90.03**
	Ameva	83.41	82.22
	Bayesian	85.23	82.41
	CACC	85.09	84.57
	CADD	83.42	83.42
	CAIM	83.59	80.91
	Chi2	88.10	90.68
	Chi-merge	85.32	87.51
	ExtChi2	80.74	82.41
	Fayyad & Irani	88.25	87.25
	PKID	80.74	81.69

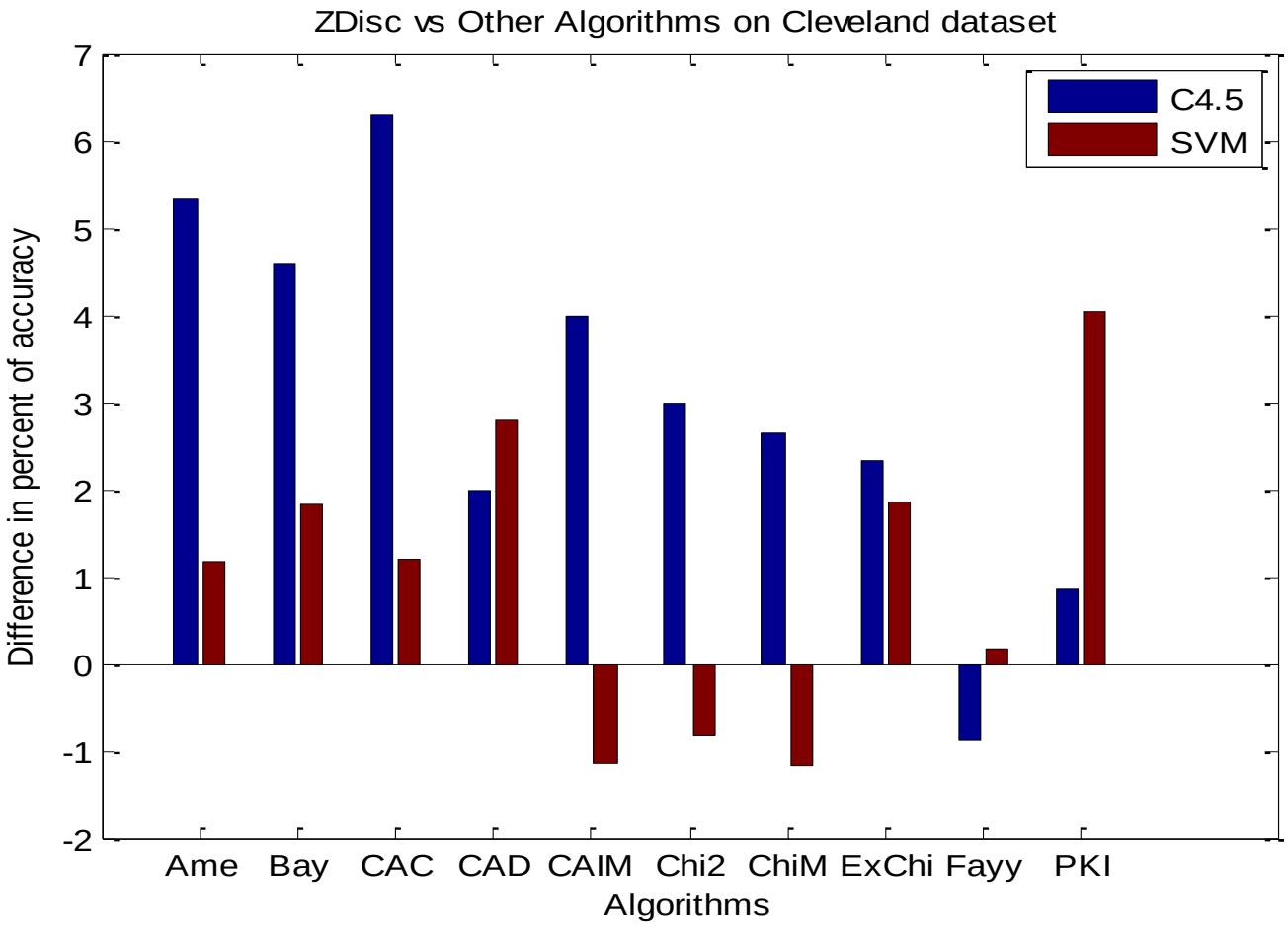
Table.5. Test classifiers of our algorithm with other discretization methods on Pima

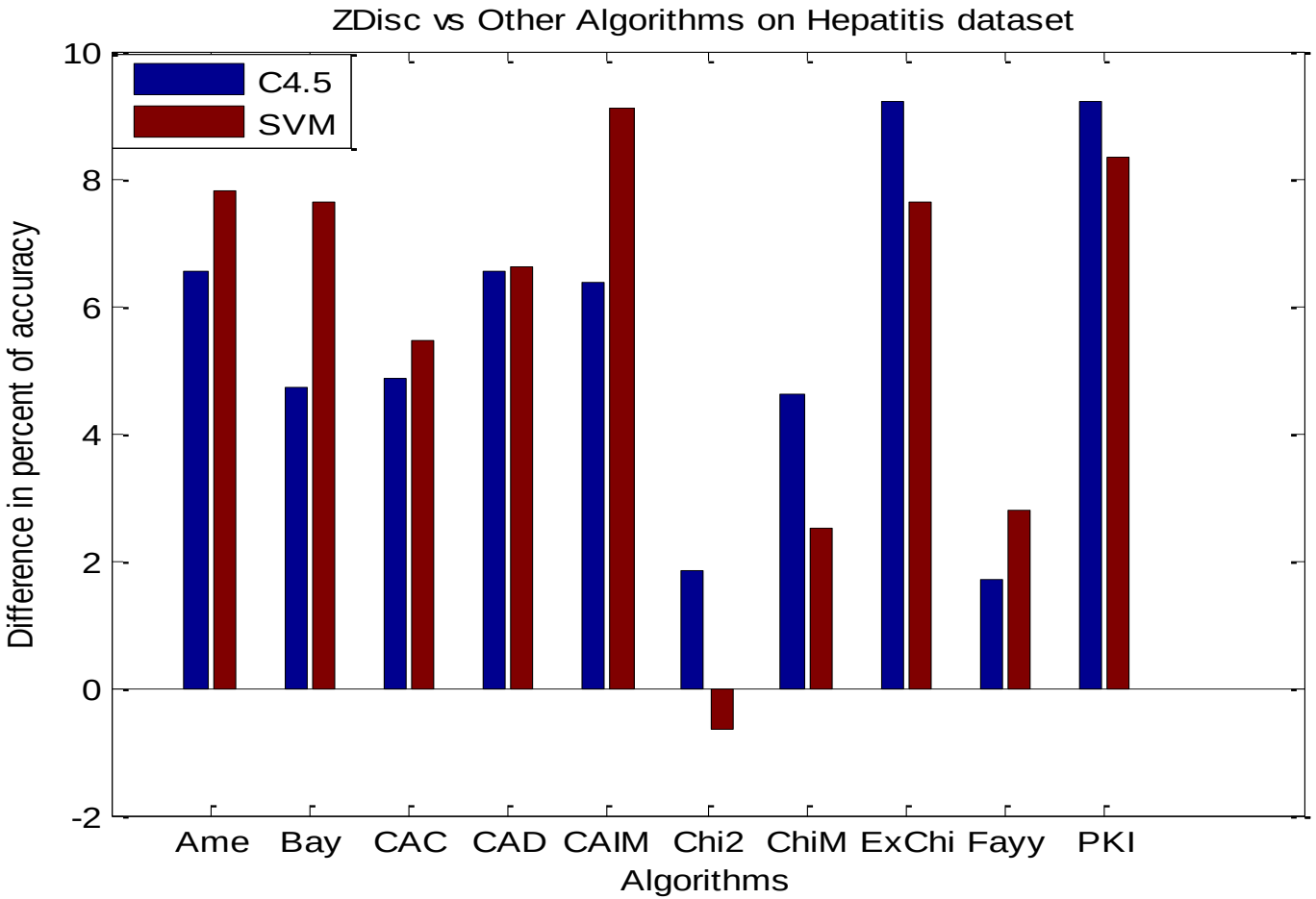
Dataset		C4.5 classifier	SVM classifier
	Discretization Algorithms	10x cross-fold Validation(%Accuracy)	10x cross-fold Validation(%Accuracy)
Pima	ZDISC	**76.17**	**76.56**
	Ameva	72.26	72.91
	Bayesian	68.01	75.66
	CACC	72.39	73.31
	CADD	65.10	65.10
	CAIM	71.86	73.71
	Chi2	75.77	77.09
	Chi-merge	73.68	72.91
	ExtChi2	73.83	72.15
	Fayyad & Irani	79.80	75.66
	PKID	74.34	65.10

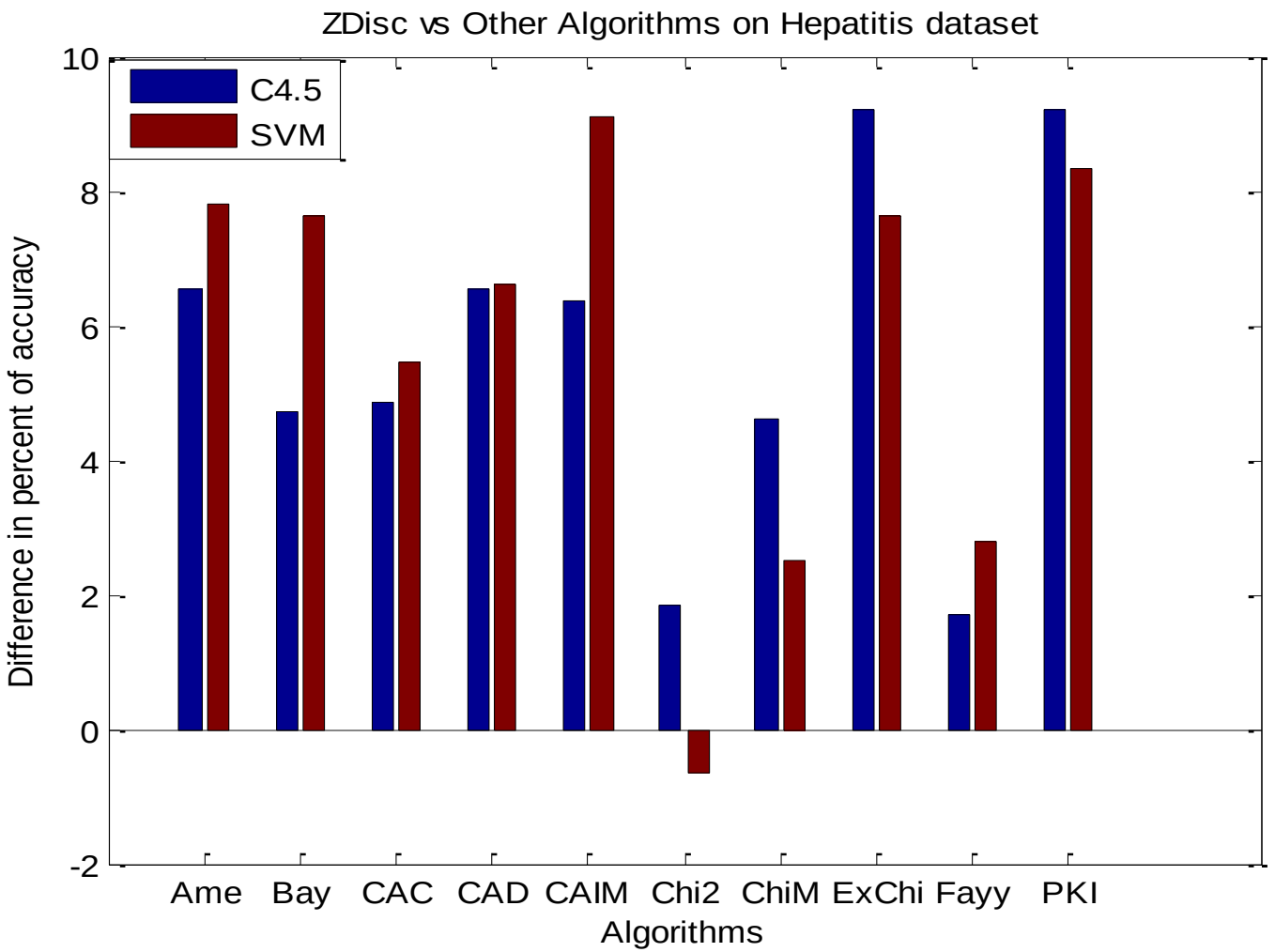
Table.6. Test classifiers of our algorithm with other discretization methods on BCW

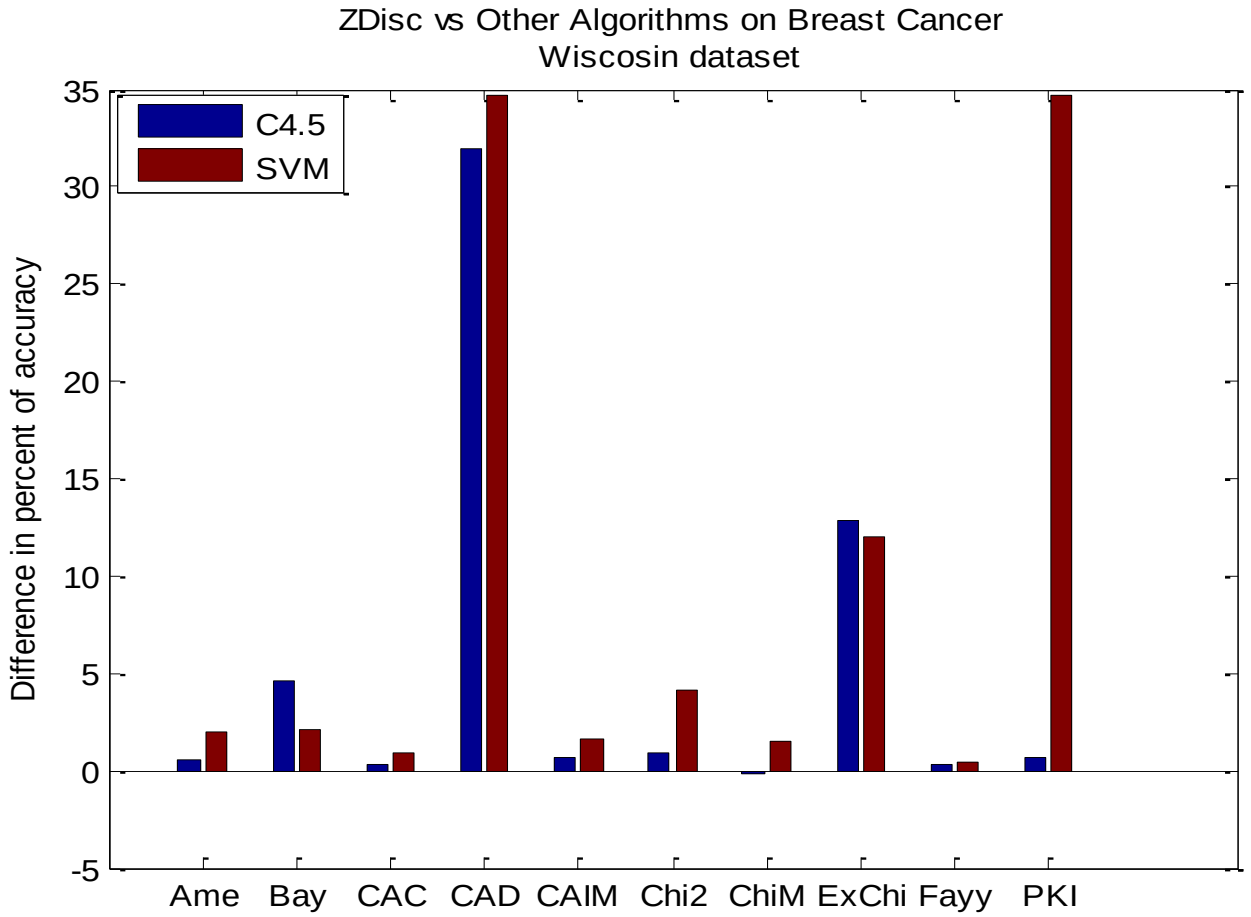
Dataset		C4.5 classifier	SVM classifier
	Discretization Algorithms	10x cross-fold Validation(%Accuracy)	10x cross-fold Validation(%Accuracy)
Breast Cancer Wiscosin	ZDISC	**94.72**	**97.41**
	Ameva	94.20	95.43
	Bayesian	90.15	95.26
	CACC	94.38	96.47
	CADD	62.74	62.74
	CAIM	94.03	95.78
	Chi2	93.85	93.32
	Chimerge	94.90	95.95
	ExtChi2	81.91	85.41
	Fayyad & Irani	94.38	97.01
	PKID	94.02	62.74


ZDisc Vs Other Algorithms on Appendicitis dataset
C4.5
SVM
(%) Differences in terms of Accuracy
8
7
6
5
4
3
2
1
0
−1
−2
AME
BAY
CAC
CAD
CAIM
Chi2
ChMe
ExCh
FAYY
PKI
Algorithms


ZDisc vs Other Algorithms on Cleveland dataset
C4.5
SVM
Difference in percent of accuracy
7
6
5
4
3
2
1
0
-1
-2
Ame
Bay
CAC
CAD
CAIM
Chi2
ChiM
ExChi
Fayy
PKI
Algorithms


ZDisc vs Other Algorithms on Hepatitis dataset
C4.5
SVM
Difference in percent of accuracy
10
8
6
4
2
0
-2
Ame
Bay
CAC
CAD
CAIM
Chi2
ChiM
ExChi
Fayy
PKI
Algorithms


ZDisc vs Other Algorithms on Hepatitis dataset
C4.5
SVM
Difference in percent of accuracy
10
8
6
4
2
0
-2
Ame
Bay
CAC
CAD
CAIM
Chi2
ChiM
ExChi
Fayy
PKI
Algorithms


ZDisc vs Other Algorithms on Breast Cancer
Wiscosin dataset
C4.5
SVM
Difference in percent of accuracy
35
30
25
20
15
10
5
0
-5
Ame
Bay
CAC
CAD
CAIM
Chi2
ChiM
ExChi
Fayy
PKI

CONCLUSIONS

- In this paper, we proposed a new discretization measure based algorithm, which aims to improve in terms of classification accuracy.
- We compared with the state-of-the art methodologies of discretization algorithms on benchmark biomedical datasets.
- The results show that a significant improvement in terms of accuracy can be achieved by applying our algorithm.
- In the future work, we will propose the fuzzy discretization index measure imputation algorithm for missing continuous values in real-world datasets.

Acknowledgement

The author would like to thank the Associates of ITQM 2014 members for their valuable support.

References

- An. A, Cercone.N, "Discretization of Continuous Attributes for Learning Classification Rules", 3rd Pacific-Asia Conference, Methodologies for Knowledge Discovery and Data Mining, 1999, pp.509-514.
- Ying Yang, Geoffrey I. Webb, and Xindong Wu, " Discretization Methods", Data Mining and Knowledge Discovery Handbook, Second Edition, O. Maimon, L. Rokach, Eds, 2010, pp. 101-116.
- M.Gethsiyal Augasta, T.Kathirvalakumar, " A New Discretization Algorithm based on Range Coefficient of Dispersion and Skewness for Neural Networks Classifier", Applied Soft Computing, vol. 12, 2012, pp.619-625.
- Dougherty J, Kohavi R, Sahami M, "Supervised and unsupervised discretization of continuous features". In Proceedings of the 12th International Conference on Machine Learning, 1995, pp.194-202.
- Kerber R , Chimerge: "Discretization for numeric attributes, In National Conference on Artificial Intelligence", AAAI Press,1992, pp.123-128.
- Kohavi R, Sahami M, "Error-based and entropy-based discretization of continuous features", In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996, pp.114-119.
- Cios, K.J and Kurgan L.A., " CLIP : Hybrid Inductive Machine Learning Algorithms that Generates Inequality Rules", Information Science, Vol.163, 2004, pp.37-83.

- Clark.P, Niblett.T, "The CN2 Algorithm", Machine Learning, vol.3, 1989, pp.261-283.
- Kurgan L.A., Cios, K.J, " CAIM Discretization Algorithm", IEEE Transaction on Knowledge and Data Engineering, vol.16, 2004, pp. 145-152.
- Tsai, C.J Lee. C.I, Yang. W.P., " A Discretization Algorithm based on Class-Attribute Contingency Coefficient", Information Sciences, vol.178, 2008, pp.714-731.
- Butterworth.R, Simovici.D.A., et al., " A Greedy Algorithm for Supervised Discretization", Biomedical Informatics, vol.37, 2004, pp.285-292.
- Fayyad.U.M, Irani. K.B., " Multi-Interval Discretization of Continuous- Valued Attributes for Classification Learning", Proceddings of 13th International Conference on Artificial Inatelligence, 1993, pp.1022-1027.
- Amitava Roy, Sankar K.Pal., " Fuzzy Discretization of Feature Space for a Rough Set Classifier", Pattern Recognition Letters, vol.24, 2003, pp.895-902.

- Soman.P, Diwakar.S, Ajay.V, "Insight into Data Mining", Prentice Hall of India, 2006.
- Holte, R.C., "Very Simple Classification Rules Reform Well on Most Commonly Used Datasets", Machine Learning, vol.11, no.1, 1993, pp.63-90.
- Liu.H, Setiono.R., "Feature Selection via Discretization", IEEE Transactions on Knowledge and Data Engineering, vol.9, 1992, pp. 642-645.
- Tay. F, Shen.L., "A Modified Chi2 Algorithm for Discretization", IEEE Transactions on Knowledge and Data Engineering, vol.14, 2002, pp. 666-670.
- Su, C.T., Hsu, J.H., " An Extended Chi2 Algorithm for Discretization of Real Value Attributes", IEEE Transactions on Knowledge and Data Engineering, vol.17, 2005, pp. 437-441.
- Wong, A.K.C and Chiu, D.K.Y., "Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data", IEEE Transactions Pattern Analysis and Machine Intelligence, vol.PAMI9, no.6, 1987, pp.786-805.

- Khurram Shehzad, "EDISC: A Class-Tailored Discretization Technique for Rule-Based Classification", IEEE Transactions on Knowledge and Data Engineering, vol.24, No.8, August 2012, pp. 1435-1447.
- Chang-Hwan Lee., "A Hellinger-based discretization method for numeric attributes in classification learning", Knowledge based Systems, vol.20, 2007,pp.419-425.
- Francisco J. Ruiz, Cecilio Angulo, and Nu´ ria Agell, "IDD: A Supervised Interval Distance-Based Method for Discretization" , IEEE Transactions Knowledge and Data Engineering, vol.20, No.9, Sept 2008, pp. 1230-1238.
- Jing, R., Breitbart, Y., "Data Discretization Unification", IEEE International Conference on Data Mining, 2007, pp.183-
- Berzal, F., et al., " Building Multi-way decision Trees with Numerical Attributes", Information Sciences vol.165, 2004, pp.73–90.
- L. Gonzalez-Abril, F.J. Cuberos, F. Velasco, J.A. Ortega. Ameva: An autonomous discretization algorithm. Expert Systems with
- Applications, vol. 36,2009,pp 5327-5332.
- X. Wu. "A Bayesian Discretizer for Real-Valued Attributes". The. Computer J. vol. 39(8), 1996, pp. 688-691.

- J.Y. Ching, A.K.C. Wong, K.C.C. Chan. "Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17(7),1995, pp. 641-651.
- Ying Yang, Geoffrey I. Webb, " Proportional k-Interval Discretization for Naive-Bayes Classifiers". 12th European Conference on Machine Learning, 2001, pp.564-575.
- K.A. Kaufman, R.S. Michalski, Learning from inconsistent and noisy data: the AQ18 approach, in: Proceeding of Eleventh International Symposium on Methodologies for Intelligent Systems, 1999.
- P. Clark, T. Niblett, The CN2 algorithm, Machine Learning, vol. 3 (4), pp. 261–283, 1989.

Your Queries Please!!!