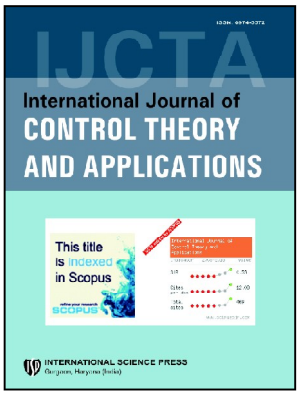

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 18 • 2017

Classification using Knowledge based Semantic Discretization

Omprakash Chandrakar[1], Jatinderkumar R. Saini[2]

[1] *Associate Professor, UkaTarsadia University, Bardoli, Gujarat, India*
[2] *Professor & I/C Director, Narmada College of Computer Applicat ion, Bharuch, Gujarat, India*

Abstract: Data can be classified into two types, Ordinal data and Nominal Data. Ordinal data can be further classified into to two categories, Continuous data and discrete data. Some machine learning techniques like decision tree, decision table, classification gives better result, when number of distinct values of the attribute is less. But in real life most of the attributes are continuous attributes, they may have infinite or very large values. In such scenario, machine learning techniques produce too many rules with relatively less accuracy. In this study, researchers used knowledge based semantically discretized Pima Indian Diabetes. This discretized dataset is used to build classification model for p redicting diabetes. Experimental result shows that accuracy of classification model built using knowledge based discretized dataset is better than the model that uses original continuous dataset. Researcher also observed, significantly less cases of false negative cases than false positive cases, which is critical for medical domain.

Keywords: Association rule mining, Data mining, Discretization, M achine learning, Pima Indian dataset.

1. INTRODUCTION

Data can be any one of the two types, Ordinal data and Nominal Data. Nominal refers to the categorically discrete data like name of city, book or type of car. Ordinal refers to quantities that have a natural ordering. Weight of a person, runs scored by batsman is some examples of ordinal values. Ordinal data can be further classified into to two categories, Continuous data and discrete data. As the name implies, continuous data have infinite number of continuous values for an attribute, while discrete data have very few values. Before apply machine learning algorithms, generally discretization is done as a pre-processing step. Discretization is the process of transforming a continuous -valued variable into a discrete one by creating a set of contiguous intervals. There are two types of discretization methods, supervised and unsupervised. Supervised discretization methods will discretize a variable to a single interval if the variable has little or no correlation with the target variable. Supervised discretization is better for classification than unsupervised d iscretization [1]. Boullé has presented a supervised discretization method [2]. Variable selection feature of discretization is also beneficial for classification [3]. Some machine learning techniques such as association rule mining [4], classification algorithms like ID3 [5] can deal with only discrete values. Even some machine learning techniques works on continuous data but they produce too many rules with less accuracy [6]. To improve the classification accuracy for diabetes

prediction, knowledge based semantically discretized Pima Indian Diabetes dataset [7] is used. Rest of the paper describes the application of classification algorithms and results are analyzed.

2. DATASET

For this study, researcher used Pima Indians Diabetes Dataset. It is taken from UCI Machine Learning Repository Data [8]. The dataset is provided by National Institute of Diabetes and Digestive and Kidney Diseases, Research Center, Applied Physics Laboratory, The Johns Hopkins University, Laurel.

2.1. Instances and Attribute Information

The Pima Indians Diabetes Dataset contains 768 records and 9 attributes including class attributes that indicates the diabetes status of a person. Attributes are as follows:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Class Distribution: Class value 1 is interpreted as "Tested positive for Diabetes"

Table 1
Class Distribution

Class Value	*Number of instances*
0	500
1	268

3. RESEARCH METHODOLOGY

In this research work, researchers studied the impact of classification accuracy on knowledge based semantically discretized data set over continuous dataset. 11 popular classification algorithms are applied on original Pima Indian Dataset and semantically discretized dataset. Classification accuracy is compared.

4. EXPERIMENTS

Tool Used: Weka 3.6:

Weka [9] [10], a data mining tool used for research purpose, is used to carry out the experiments in this study.

Semantically Discretized Pima Indian Diabetes Dataset:

Chandrakar and Saini proposed knowledge based semantic discretization process and applied it on Pima Indian Diabetes dataset [11]. Table 2 presents the most appropriate discretization rules they found.

Table 2
Discretization Rules and their corresponding Confidence

No	*Attribute*	*Discretization Rule*	*Confidence*	
			If RISK = "High" then Diabetes = "Yes"	*If RISK = "Low" then Diabetes = "No"*
1	Number of times pregnant	ifelse(A<5,0,1)	48%	72%
2	Plasma glucoseconcentration a 2 hours in an oral glucose tolerance test	ifelse(A<6,0,1)	51%	71%
3	Diastolic blood pressure(mm Hg)	ifelse(A<7,0,1)	56%	71%
4	Triceps skin fold thickness (mm)	ifelse(A<25, 0, 1)	-	70%
5	2-Hour serum insulin (mu U/ml)	ifelse(A<75, 0, 1)	-	70%
6	Body mass index (weightin kg/(height in m)^2)	ifelse(A<30, 0, 1)	-	83
7	Diabetes pedigree function	ifelse(A<0.35, 0, 1)	41%	72%
8	Age (years)	ifelse(A<35, 0, 1)	49%	79%

Using the above rules, semantic discretization is performed on the Pima Indian Dataset. 11 common classification algorithms are applied on original Pima Indian Dataset and semantically discretized dataset.

Researchers not found significant semantic rule for two attributes, Diastolic blood pressure (mm Hg) and Triceps skin fold thickness (mm). Researchers envision that since there is no significant rule, they might not be contributing significantly in classification. So we performed the whole experiment twice, including all attributes and excluding the above two attributes.

Table 3 shows classification accuracy with and without Semantic Discretization. 11 classification algorithm mentioned in table 4, are performed on following dataset.

Table 3
Classification Accuracy with and without Semantic Discretization

No	*Algorithm*	*Prediction Accuracy*			
		Including all attributes		*Excluding Attribute 3 & 4*	
		Original Dataset	*Semantically Discretized Dataset*	*Original Dataset*	*Semantically Discretized Dataset*
1.	*Bayes Logistic Regression	66.012	NA	63.802	NA
2.	Navies Base	76.302	76.302	75.912	76.432
3.	Bayes Net	74.349	76.302	74.349	76.432
4.	PART	74.479	73.828	74.479	77.745
5.	AD Tree	72.917	75.781	73.958	75.651
6.	BF Tree	73.568	74.870	73.307	77.747
7.	Decision Stump	71.875	72.266	71.875	72.266
8.	J 48	73.828	75.130	74.349	75.391
9.	NB Tree	74.349	76.432	74.740	76.432
10.	*ID3	NA	70.703	NA	76.432
11.	*VAODE	NA	74.740	NA	76.823

*Some algorithms support only discrete or continuous dataset.

1. Original Dataset including all attributes
2. Semantically Discretized Dataset including all attributes
3. Original Dataset excluding attributes Diastolic blood pressure (mm Hg) and Triceps skin fold thickness (mm)
4. Semantically Discretized Dataset excluding attributes Diastolic blood pressure (mm Hg) and Triceps skin fold thickn ess (mm)

5. RESULT ANALYSIS

1. Table 3 shows that out of 19 experiments, classification accuracy either increased in 53% cases or remains same in 32%, after excluding two attributes from the dataset. It decreases in only in 15% cases. So it can be concluded that the two attributes Diastolic blood pressure (mm Hg) and Triceps skin fold thickness (mm) can be excluded.
2. After considering the dataset excluding two attributes, compare the classification accuracy with the original dataset and semantically discretized dataset. There is total 7 experiments, and found the accuracy increased in all cases.
3. Considering the dataset with all, researchers compare the classification accuracy with the original dataset and semantically discretized dataset. There is total 8 experiments, and found the accuracy increased in 6 cases, remains same in 1 case and decreased in only 1 case.

Table 4, summarizes how the classification accuracy increases/decreases on Semantically Discretized Dataset over Original Dataset.

Table 4
Classification Accuracy increases/decreases with and without Semantic Discretization

	Total No of Exp.	*Accuracy Increased*	*Accuracy Remains Same*	*Accuracy Decreased*
Dataset excluding 2 attributes	7	7	0	0
Dataset with allattributes	8	6	1	1

6. CONCLUSION

From the table 4, researchers found that 14 out of 15 experiments, classification accuracy increased when dataset is semantically discretized. So it can be concluded that discretizing the data semantically using association ru le mining increases the accuracy of classification significantly.

REFERENCES

[1] Kohavi R, Sahami M . Error-based and entropy -based discretization of continuous features. In: Proceedings of the Second International Conference on Knowledge Discovery and Data M ining; 1996; Portland, Oregon: AAAI Press; 1996. p. 114-119.

[2] Liu H, Setiono R. Feature selection via discretization. Knowledge and Data Engineering 1997;9(4):642-645.

[3] Boullé M . M odl: A bayes optimal discretization method for continuous attributes. M achine Learning 2006;65(1):131-165.

[4] R. Agrawal, T. Imielinski, and A. Swami. M ining association rules between sets of items in large databases. In ACM SIGM OD Conf. M anagement of Data, M ay 1993.

[5] Quinlan, J. R. 1986. Induction of Decision Trees. M ach. Learn. 1, 1 (M ar. 1986), 81-106

[6] Improving Classification Performance with Discretization on Biomedical Datasets

[7] http://ftp.ics.uci.edu/pub/ml-repos/machine-learning databases/pima-indians-diabetes, 2003.

[8] http://archive.ics.uci.edu/ml/

[9] Ian H. Witten; Eibe Frank; M ark A. Hall (2011). "Data M ining: Practical machine learning tools and techniques, 3rd Edition". M organ Kaufmann, San Francisco. Retrieved 2011-01- 19.

[10] www.cs.waikato.ac.nz/ml/weka/

[11] Omprakash Chandrakar, Dr. Jatinderkumar R. Saini, "Knowledge based Semantic Discretization using Data M ining Techniques" Submitted to International Journal of Advanced Intelligence Paradigms