ISSN No:-2456 –2165

Performance Improvement of Classifier Using Attribute Selection with Association Rule Mining Technique

Rutuja Shinde [1], Saee Joshi [1], Aishwarya Gunjal [1], Dr. K. Rajeswari [2]
[1]P.G. Student, Department of Computer Engineering, PC College of Engineering , Akurdi, Pune, India.
[2]H.O.D, Department of Computer Engineering, P C College of Engineering, Akurdi, Pune, India.

Abstract:-Data mining refers to extracting knowledge from huge amount of data. Apriori algorithm is the paradigmatic algorithm of association rule, which enumerates all of the frequent item sets. When this algorithm encounters dense but noisy data, large number of extensible patterns emerge and hence, the algorithm's performance diminishes dramatically. In order to find more worth rules. this paper explains the radicals of Association Rule Mining (ARM)and moreover acquire a general framework. This paper proposes selection of best attributes from the best rules obtained with Apriori algorithm. After that truly explore their influences and carry forward several run time demonstrations using a classification technique Iterative Dichotomiser (ID3). The research describes algorithmic discussion and comparison of the performance of classification algorithm with and without best attributes.

Keywords:- Data Mining, Association Rules, ID3, Apriori Algorithm, Frequent item, Performance.

I. INTRODUCTION

Data mining is the process of digging out interesting facts and statistics for references and analysis from large amount of data. Data mining includes a systematic task of acquiring previously unknown facts and statistic such as deviations and several data records from large repositories of data. Multi-Formality can be achieved for various datasets by preprocessing data and the quality of data can be improved in terms of accuracy.

Association rules mining is used to locate the associations and relations among item sets of huge data. Association rules mining is of great significant branch of data mining research, and association rules is the most obvious style of data mining[5].Presently, association rules mining problems are idolized by the researchers in database, artificial intelligence, statistic, information gaining, visible, information science, and many other fields. Many extraordinary results have been found out. What can productively catch the important relationships among data are straightforward forms of association rules and easy to explain and be understandable[7].Mining association rules from large database has become the most fully fledged, precise, and active research area. Best attribute selection is essential step because it reduces the time consumed by researcher to classify the data and improve the performance[15][16][17].The best attributes are selected on the basis of rules generated in association rule mining.

The Apriori algorithm is a well known and a long established algorithm in data mining. The main idea of this approximation is to find a most prominent pattern in various sets of data. The algorithm suffers from many pitfalls. This paper deals with the apriori algorithm, and various techniques that were proposed to analyze the apriori algorithm and its types. Association Rule Mining has attracted a lot of objectives in research area of Data Mining and generation of association rules is completely relies on finding Frequent Item sets.

Classification plays an important role as a technique of data mining in the area of e-learning[6]. Classification is a predictive data mining approach, that forecast about values of data using known results found from different data. Predictive models focus on allowing to predict the unknown values of variables of interest given known values of other variables[13].Predictive modeling can be thought of as learning to best predict the probability of an outcome that depict from an input set of vector measurements to a scalar output. Classification draw data into predefined groups of classes. It is often defined as supervised learning because the classes are known before examining the data.

The ID3 algorithm is a classification algorithm that uses Information Entropy, its primary idea is that all examples are mapped to different categories following to different values of the condition attribute set; its fundamental is to decide the best classification attribute form condition attribute sets[3][10].The algorithm chooses information gain as attribute selection criteria; generally the attribute that has the highest information gain is selected as the splitting attribute of current node, in order to build information entropy that the divided subsets need smallest. As stated by the different values of the attribute, branches can be fixed, and the process above is recursively called on each branch to create other nodes and branches until all the samples in a branch belong to the same

group. To select the splitting attributes, the notion of Entropy and Information Gain are used.

II. RELATED WORK

Jiao in his paper [1], proposes an enhanced algorithm to find association rules, instead of the traditional Apriori algorithm. Finally, the enhanced algorithm is substantiate, the results show that the enhanced algorithm is reasonable and productive, can retrieve more valuable information. This study concentrates on how to solve the precise problems of Apriori algorithm and elevate another association rules mining algorithm. This paper solves the complication of data redemption and information unavailability. It hopes to extract more useful information.

Jochen et al. in their paper [2], discusses the algorithmic features of association rule mining. In fact, a wide variety of precise algorithms to extract association rules have been established during the last years. These approximations are more or less outlines separately in the coinciding literature. To beat this situation they give a typical survey of the primary ideas behind association rule mining to identify the fundamental approaches and describe them in detail. The resulting framework is used to standardize and present common approaches in ambience. Furthermore they show the common postulates and differences between the algorithms. Finally overview is concerned with a comparison of the algorithms on gaining efficiency.

Kalpesh et al. in their paper[3], developed a system which can forecast the performance of students from their previous performances using abstraction of data mining approach below Classification. The data set containing information about students, such as marks and rank in entrance examinations, gender, marks scored in the board examinations of classes X and XII and results in first year of the previous batch of students. By using the ID3 (Iterative Dichotomiser 3) and C4.5 classification algorithms on this data[14].

Deepali et al. in their paper [4], includes a detailed overview of apriori algorithm and recent achievements done in the area of apriori algorithm. With the research on various improved algorithms, it is summarized that the main concern is to generate less candidate sets which contains frequent items within a reasonable amount of time. Also, in future some more algorithms can be developed that requires only single scan for the database and are efficient for large databases. For large datasets, it saves time and cost and increases the efficiency.

K. Rajeswari in her paper[15], proposes a novel method for attribute selection using association rule mining. This technique reduces the generated rule based on desired class label attribute. This method is very productive for real datasets.

The papers [16][17], discusses about feature selection and classification technique to enhance the performance of classifier. Accuracy is improved by dimensionality reduction as well as this methods are productive and cost effective.

III. PROPOSED METHOD

The following fig 3.1 shows overall comparison in terms of accuracy by using ID3 and Apriori algorithm.

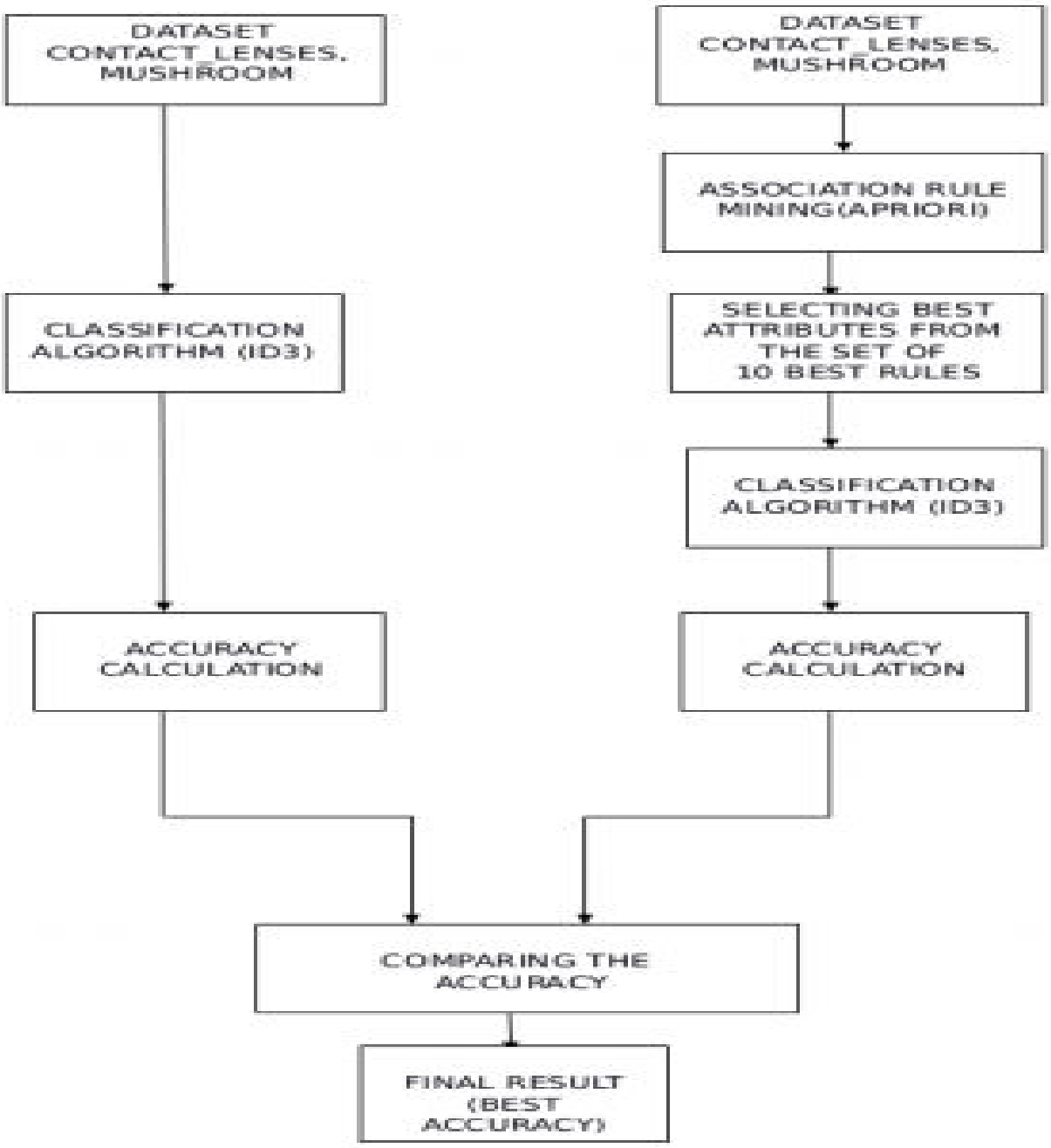


Fig 1: Work Methodology

A. Dataset

Here basically two datasets used namely contact-lenses and Mushrooms which are taken from weka tool[8][9]. Contact-lenses dataset has 5 attributes and 24 instances. Mushroom dataset has 23 attributes and 8124 instances.

B. Apply Apriori algorithm[10]

Apriori algorithm is applied on both the datasets. and used all attributes for finding best rules. Here found best 10 rules for both the datasets.The core principles of this algorithm are the subsets of frequent itemsets are frequent itemsets and the supersets of infrequent item sets are infrequent item sets. An algorithm decreases the number of candidate items in the candidate item set to find the best attribute.

C. Apply ID3 algorithm[10]

Apply ID3 algorithm on both dataset and on all attributes. And check the accuracy of datasets. In decision tree learning, ID3 (Iterative Dichotomiser (3) is an algorithm invented by Ross

ISSN No:-2456 –2165

Quinlan used to generate a decision tree from the dataset[11].The decision tree process involves constructing a tree to model the classification process. In ID3 calculate the entropy and information gain to find the splitting criteria for classification.

D. Preprocess Data[10]

As removal of irrelevant, noisy attributes consumes 85% of researcher's time, this paper has used ARM to select best attributes from the best rules obtained from step 2.

E. Re Apply ID3 algorithm

After selecting best attribute again apply ID3 on dataset and check the accuracy obtain.

F. Compare Accuracy and note findings

Compare accuracy obtained in step 3 and step 5. it has been observed that the accuracy of both datasets are increased. Accuracy is measured in terms of number of attributes are correctly classified.

IV. DATA SET

A. UCI Data Set [8]

The University of California Irvine (UCI) is a expert system to achieve different datasets. David Aha and associate graduate students at UC Irvine created as an ftp archive in 1987.It consists of 404 datasets as facility for intelligent retrieval.UCI consists of default task as classification, regression, clustering and others. Datasets can be classified into three different types of attributes such as Numerical, Categorical and Mixed. This expert systems consist of different types of datasets which fall under various fields of research like Life Science, Physical Science, CS /Engineering, Social Science, Business, Game and other. This paper uses two datasets to check the performance of association rule mining.

B. Contact-Lenses Dataset[8].

From weka tool have been used to evaluate the performance of apriori and ID3. The characteristics of dataset composed of nominal attributes as shown in the table below.

Relation: contact-lenses

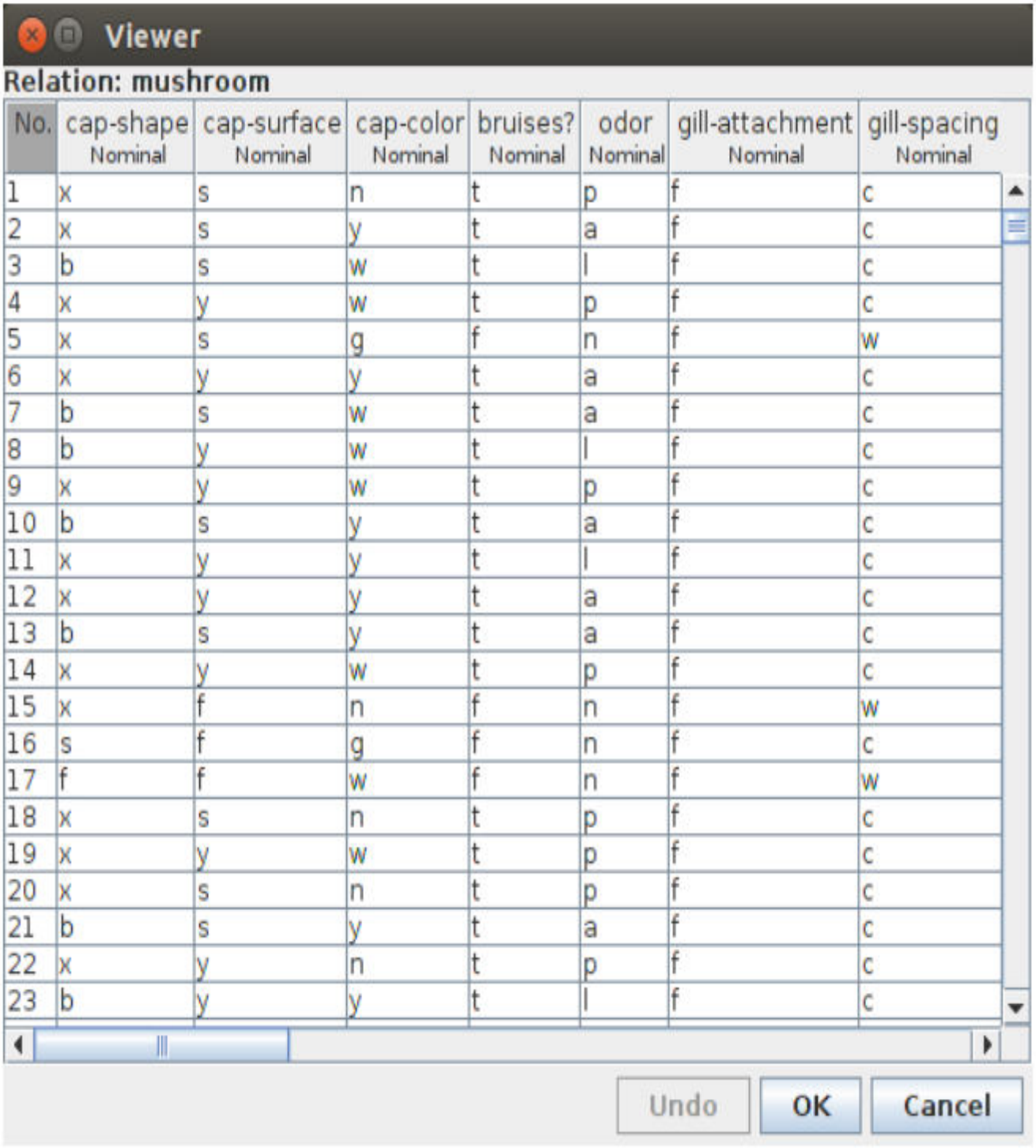
No.	age Nominal	spectacle-prescrip Nominal	astigmatism Nominal	tear-prod-rate Nominal	contact Nom
1	young	myope	no	reduced	none
2	young	myope	no	normal	soft
3	young	myope	yes	reduced	none
4	young	myope	yes	normal	hard
5	young	hypermetrope	no	reduced	none
6	young	hypermetrope	no	normal	soft
7	young	hypermetrope	yes	reduced	none
8	young	hypermetrope	yes	normal	hard
9	pre-...	myope	no	reduced	none
10	pre-...	myope	no	normal	soft
11	pre-...	myope	yes	reduced	none
12	pre-...	myope	yes	normal	hard
13	pre-...	hypermetrope	no	reduced	none
14	pre-...	hypermetrope	no	normal	soft
15	pre-...	hypermetrope	yes	reduced	none
16	pre-...	hypermetrope	yes	normal	none
17	pres...	myope	no	reduced	none
18	pres...	myope	no	normal	none
19	pres...	myope	yes	reduced	none
20	pres...	myope	yes	normal	hard
21	pres...	hypermetrope	no	reduced	none
22	pres...	hypermetrope	no	normal	soft
23	pres...	hypermetrope	yes	reduced	none

Table 1: Contact Lens Dataset[8]

The above table shows that there are two kinds of attribute; categorical are in the set type as the attributes(age, spectacle-prescript, tear-prod-rate, contact-lenses) and binary categories are all the attributes that represent as yes/no or True/False in their classes; attributes(astigmatism).

C. Mushroom Dataset[9]

From weka tool have been used to evaluate the performance of apriori and ID3. The characteristics of dataset composed of nominal attributes as shown in the table below.



No.	cap-shape Nominal	cap-surface Nominal	cap-color Nominal	bruises? Nominal	odor Nominal	gill-attachment Nominal	gill-spacing Nominal
1	x	s	n	t	p	f	c
2	x	s	y	t	a	f	c
3	b	s	w	t	l	f	c
4	x	y	w	t	p	f	c
5	x	s	g	f	n	f	w
6	x	y	y	t	a	f	c
7	b	s	w	t	a	f	c
8	b	y	w	t	l	f	c
9	x	y	w	t	p	f	c
10	b	s	y	t	a	f	c
11	x	y	y	t	l	f	c
12	x	y	y	t	a	f	c
13	b	s	y	t	a	f	c
14	x	y	w	t	p	f	c
15	x	f	n	f	n	f	w
16	s	f	g	f	n	f	c
17	f	f	w	f	n	f	w
18	x	s	n	t	p	f	c
19	x	y	w	t	p	f	c
20	x	s	n	t	p	f	c
21	b	s	y	t	a	f	c
22	x	y	n	t	p	f	c
23	b	y	y	t	l	f	c

Table 2: Mushroom Dataset[9]

Above table have total 22 attribute such ascap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gillspacing ,gill-size ,gill-color, stalk-shape, ring-number stalk-root, stalk-surface-above-ring ,stalk-surface-below-ring, stalk-color-above-ring , stalk-color-below-ring ,veil-type, veil-color, ,ring-type, spore-print-color, population, habitat.

V. RESULTS AND DISCUSSION

A. Weka Tool [8]

Weka is abbreviation of Waikato Environment for Knowledge Analysis Tool. It is a machine learning software that developed using java programming language. Weka is open source platform for analyzing data and predictive modeling. It supports portability thus run on many computing platform. Weka consists of several data mining task such as Data Preprocessing, Regression, Feature Selection, Classification, Clustering and Visualization.

B. Performance Measures

A data mining practitioner has a distinct aim to choose a feature selection method that increase their chances of having the optimal metric for their single dataset of interest. The performance of classifier is measure using Weka 3.6.10 tool.

A) Accuracy[10]

Accuracy is measured in terms of correctly classified tulles. It is calculated and represented in percentiles.

$$\text{Accuracy} - \frac{TP+FP}{P+N}$$

where, TP= True positive

FP=False positive

P=TP+FP N=TN+FN

C. Results and Analysis

No.	Datasets	Preliminary Attributes	Accuracy using ID3	Best attributes using Association Rule Mining	Accuracy using ID3
1.	Contact_lenses	5	70%	4	83%
2.	Mushroom	23	96%	3	98%

Table 3: Accuracy Measure

From the above table it can be observed that the accuracy of contact-lenses after applying ID3 is 70% and mushroom dataset is 96%. After removing noisy attributes and gaining the best attribute when ID3 algorithm was applied again than the accuracy had increased for contact lenses dataset by 83% and for mushroom dataset by 98% respectively.

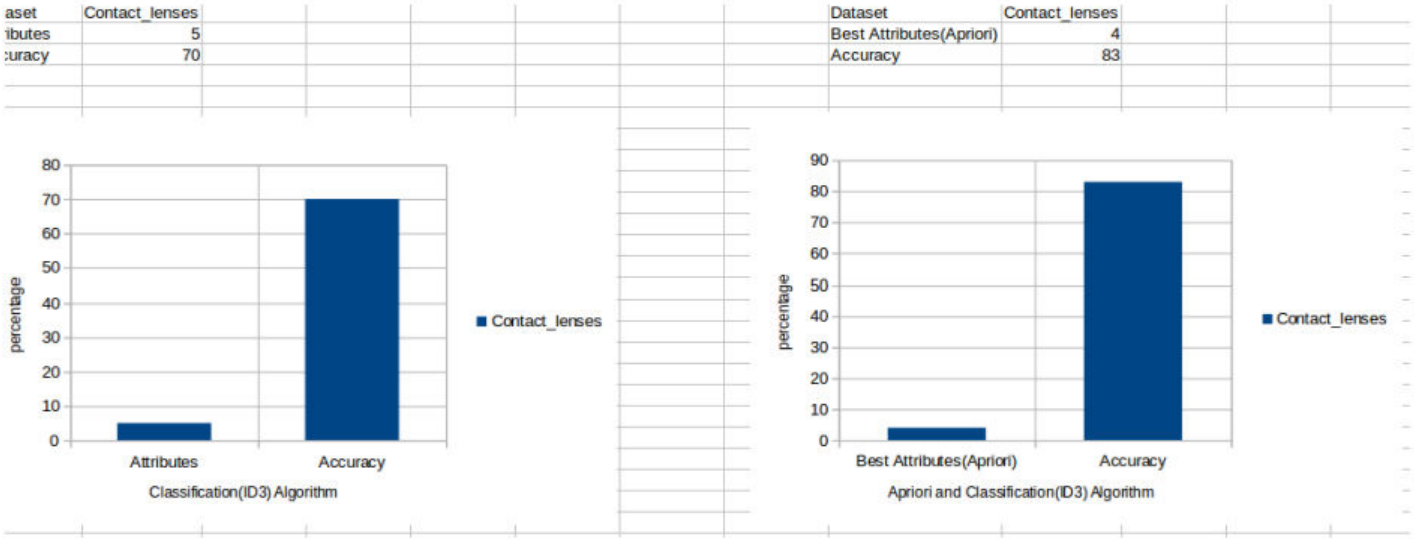


Fig 2: Comparative Analysis for the Dataset Contact-Lenses

ISSN No:-2456 –2165

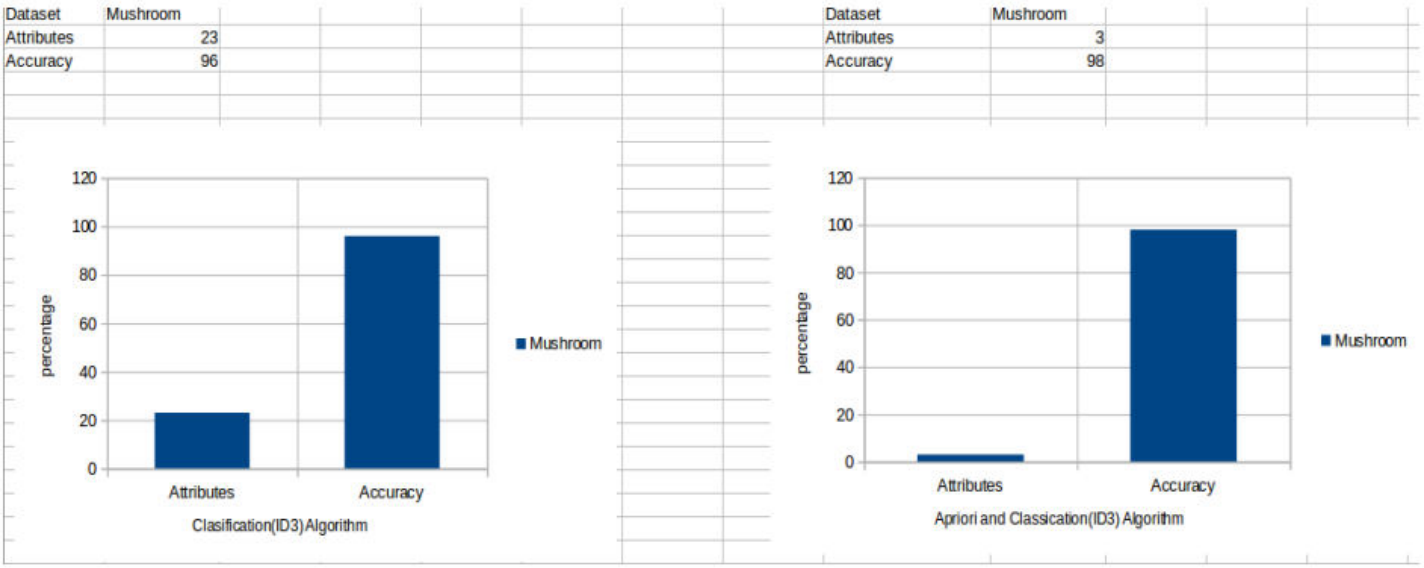


Fig 3: Comparative Analysis for the Dataset Mushroom

VI. CONCLUSION

This paper proposes the method of selecting best attribute from the best rules optimized using apriori algorithm. It is observed that the performance in terms of accuracy has been significantly increased. The results show that the method applying classification technique only on the best attributes has proven to be helpful in achieving improved accuracy.

REFERENCES

[1]. Jiao Yabing,Research of an Improved Apriori Algorithm in Data Mining Association Rules,International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.
[2]. Jochen Hipp,Ulrich G untzer,Gholamreza Nakhaeizadeh,Algorithms for Association Rule Mining – A General Survey and Comparison,July 2000.
[3]. Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, predicting students' performance using id3 and c4.5 classification algorithms,2013.
[4]. Deepali Bhende , Ushakosarker , MnishaGedam,Study of various Improved Apriori Algorithms,National Conference on Recent Trends in Computer Science and Information Technology, 2016.
[5]. K. S. Ranjith, Yang Zhenning, Ronnie D. Caytiles ,N. Ch. S. N. Iyengar,Comparative Analysis of Association Rule Mining Algorithms for the Distributed Data,International Journal of Advanced Science and Technology Vol.102 (2017).
[6]. Brijesh Kumar Bhardwaj, Saurabh Pal, Data Mining: A prediction for performance improvement using classification, 2011.
[7]. K.Saravana Kumar, R.Manicka Chezian, A Survey on Association Rule Mining using Apriori Algorithm, Volume 45– No.5, May 2012.
[8]. https://storm.cis.fordham.edu/~gweiss/datamining/datasets.htm.
[9]. https://archive.ics.uci.edu/ml/datasets/Mushroom.
[10]. Han,Jiawei Kamber, Micheline Pei and Jian, "Data Mining: Concepts and Techniques" Elevier Publishers Third Edition, ISBN: 9780123814791, 9780123814807.
[11]. J.r. quinlan, Induction of Decision Trees, J. Machine Learning 1: 81-106, 1986.
[12]. Pierre Baldi ,SorenBrunak ,Yves Chauvin ,Claus A. F. Andersen, Henrik Nielsen , Assessing the accuracy of prediction algorithms for classification: an overview ,Vol 16 no 5 2000.
[13]. Surjeet Kumar Yadav, Saurabh Pal , Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, Vol. 2, No. 2, 51-56, 2012.
[14]. Anuja Priyama , Abhijeeta , Rahul Gupta , Anju Ratheeb , Saurabh Srivastava , Comparative Analysis of Decision Tree Classification Algorithms , Vol.3, No.2 (June 2013).
[15]. K. Rajeswari ,Feature Selection by Mining Optimized Association Rules based on Apriori Algorithm, Volume 119 – No.20, June 2015.
[16]. Nuntawut Kaoungku, Keerachart Suksut, Ratiporn Chanklan, Kittisak Kerdprasop, Nittaya Kerdprasop, Data Classification Based on Feature Selection with Association Rule Mining, March 15 - 17, 2017.
[17]. K.Rajeswari , V.Vaithiyanathan ,Shailaja V.Pede, Feature Selection for Classification in Medical Data Mining, Volume 2, Issue 2, March – April 2013.