JMIM: A Feature Selection Technique using Joint Mutual Information along with the Maximum of Minimum Approach

Rajlakshmi S. Saner[1] Dr. S. S. Sane[2]
[1]Research Student [2]Vice Principle & HOD
[1,2]Department of Computer Engineering
[1,2]K.K.Wagh College Of Engineering- Nashik, Savitribai Phule University of Pune, Maharashtra, India

Abstract— Feature selection using information theory increases the computational efficiency, scalability in terms of the dataset dimensionality irrespective of any classifier. However, researchers have observed some drawbacks. One of those is the amount of redundancy in the generated results. This redundancy increases with the increase in relevance. To overcome this limitation, JMI (Joint Mutual Information) technique has been proposed. This algorithm takes into account both - the relevance and redundancy along with the class label while calculating the mutual information. In this research work the technique named as Joint Mutual Information Maximization - JMIM is presented. This algorithm combines the JMI along with the 'maximum of minimum' approach to overcome the limitation of redundancy. In Experimental Analysis, a comparative study of JMIM feature set and WEKA feature selection technique is presented. Three classifiers have been used to test the accuracy of feature set generated by JMIM and WEKA. It investigates the performance of the algorithm in comparison with the 'Attribute Selection' technique from WEKA. Experimental results show that the JMIM algorithm gives better performance as compared to Attribute Selection. But, the number of features required to generate these results is much higher as compared to Attribute Selection.

Key words: Mutual Information, feature selection, classification, joint mutual information

I. INTRODUCTION

The process of selecting features is also known as, "Variable Subset Selection" or "Attribute Selection". Feature Selection helps to minimize number of attributes from given dataset and also to simplify the model interpretation. It is a technique used in context where there are many features and data points are contained by dataset. It returns the subset of features. In the study of Supervised and unsupervised learning high dimensional data is significant problem. Hence the data dimensionality is decreased by keeping the feature set of the data as low as possible, which also decreases the training time and enhances the accuracy of the dataset. Feature extraction and feature selection are two different methods used to reduce the dimensionality.

To work with large data set and identification of the relevant features among the multiple attribute has become one of the critical tasks in the field of data mining. With the advancements in technology, researchers have developed many algorithms to overpower the previous algorithms/techniques.

This process of selecting features also helps to minimize number of attributes from given dataset and also to simplify the model interpretation. It is a technique used in context where there are many features and data points are contained by dataset.

The basic aim of feature selection is: [2, 3]

1) To minimize generalization error and create best smallest subset of k-features.
2) To improve generalization performance as compared to the whole model built without spec-
3) To improve generalization performance as compared to the whole model built without specific feature set.
4) Create a smaller, simpler dataset for better understanding of the entire filtered data generation.

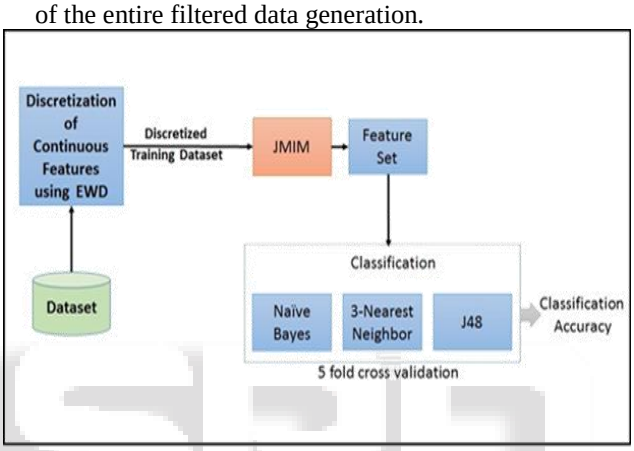

Fig. 1: System Architecture

Feature selection strategy is nothing but a pre-processing step or it can be used in conjunction with machine learning algorithms used for classification and regression purpose.

II. OVERVIEW

This work proposes a feature selection mechanism over high dimensional dataset. The Proposed system introduces feature selection mechanisms known as Joint Mutual Information Maximization to achieve feature selection accuracy. The selected subset has high mutual information score with class label C. On selected feature set Naive Bayes, 3-Nearest Neighbour and J48 algorithms are applied for classification and accuracy evaluation. The selected feature set provides near or equal accuracy of classification as compared with original dataset.

Data Discretization: In Data Discretization stage, the continuous features are discretized into nominal values depending upon the number of class labels. The EWD (Equal Width Discretization) divides the nth attributes of every instance using the below given technique: Firstly, the mid of the nth attribute of every instance is calculated: (max-min)/2.

If all the values of the nth attribute are less than the mid value of all the attributes, the attribute is termed as 'All' and all the attributes are assigned the nominal value 0.

If in case the values differ or are higher than the mid value of all attributes, the ranges are defined. If the value of the attribute is higher than the mid value, the

nominal value 1 is assigned, else the nominal value 0 is assigned to the attribute.

	A 1	A 2	A 3	A 4	A 5	C
I1	0.0200	0.0371	0.0128	0.0207	0.0954	R
I2	0.0453	0.0523	0.0243	0.0689	***0.1183***	R
I3	0.0262	0.0539	0.0299	0.1083	0.0974	M
I4	***0.0100***	0.0171	0.0423	0.0205	0.0205	R

Table 1: Real Values from Sonar Dataset.

Centroid = [Max (0.1183) – Min (0.0100)]/2 = (0.1081)/2 =0.0540

MI Score for every attribute is evaluated. For every iteration, one attribute is selected. This attribute is added to the feature set, and is eliminated from the feature list.

	A 1	A 2	A 3	A 4	A 5	C
I1	'All'	'All'	'All'	-inf-0.0439	0.0490-inf	R
I2	'All'	'All'	'All'	0.0440-inf	0.0490-inf	R
I3	'All'	'All'	'All'	0.0440-inf	0.0490-inf	M
I4	'All'	'All'	'All'	-inf-0.0439	-inf-0.0489	R

Table 2: Labeling according to centroid & mid-values

	A1	A2	A3	A4	A5	C
I1	0	0	0	0	1	R
I2	0	0	0	1	1	R
I3	0	0	0	1	1	M
I4	0	0	0	0	0	R

Table 3: Discretized Dataset

III. ALGORITHM

Input = D: Dataset
Output = Feature Set, Classification Accuracy
Processing:

1) Read Text file 'F'.
2) Convert F into ARFF format.
3) Apply discretization.
4) Get top-ranked feature 'f' by calculating Mutual Information Score with the class value.
5) Add class value and the top-ranked feature 'f' in the feature set.
6) While (true) {
   for all attributes a in I (I = all instances)
   set tmi = 0 {
   t1 = get attributes a
   for all the features in f-list {
   t2 = feature in f-list
   tmi + = calculate MI (t1 , t2) }
   add tmi to table with attribute }
   sort using tmi value (table values)
   get first attribute from table a1
   add a1 in the feature set
   check feature list.size + 1 = = N
   break }
   End while.
7) Extract N features
8) Create ARFF file of 'n' features
9) Save the ARFF file and display
10) Carry out classification using WEKA libraries on the feature set – Naïve Bayes, 3-Nearest Neighbor and J48 Classifiers.
11) Display Accuracy in % and save the ARFF files.
12) End.

IV. EXPERIMENTAL SETUP

Dataset	Attributes	Instances	Classes
Sonar	61	208	2
Libra	91	360	15
Ionosphere	35	351	2
Credit G	21	1000	2
Segment Challenge	20	1500	7
Glass	10	214	6

Table 4: Dataset Description

In the experiments, 6 benchmark datasets are used to test the performance of the algorithm. Some of these datasets are downloaded from the UCI Repository whereas some are used from the WEKA data directory. These datasets includes Sonar, Libra Movement, Ionosphere, Credit-G, Segment Challenge, and Glass Datasets. The information of these datasets is listed in above table 4. In experiments, the JMIM algorithm is compared with the feature selection technique used in WEKA. Experiments were performed in i3 processor and on Windows-7 operating system. Implementation of JMIM along with the 3 classifiers is done on NetBeans IDE 8.0.1. The experiments were perform-med on k=3 and 5-fold cross validation as the performance metric for classification accuracy. Following graphs give the performance analysis over different datasets.

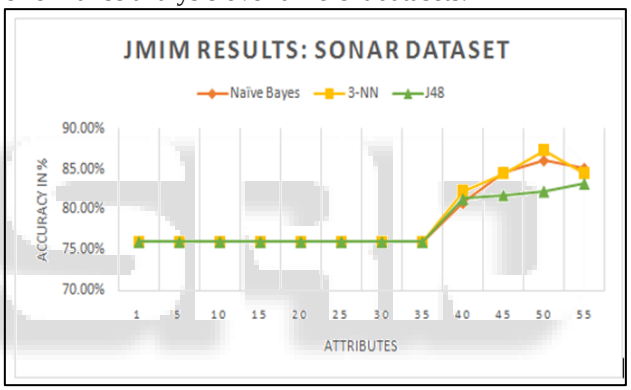

Fig. 2: JMIM results on Sonar Dataset

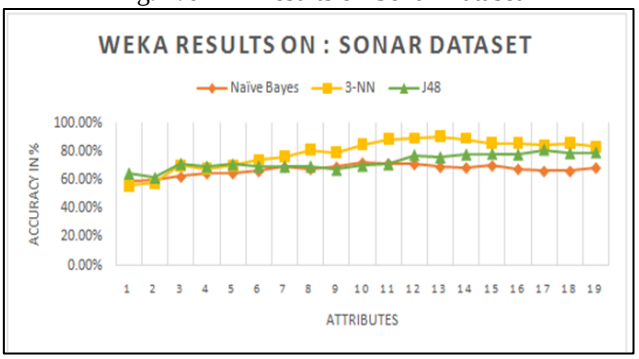

Fig. 3: WEKA results on Sonar Dataset

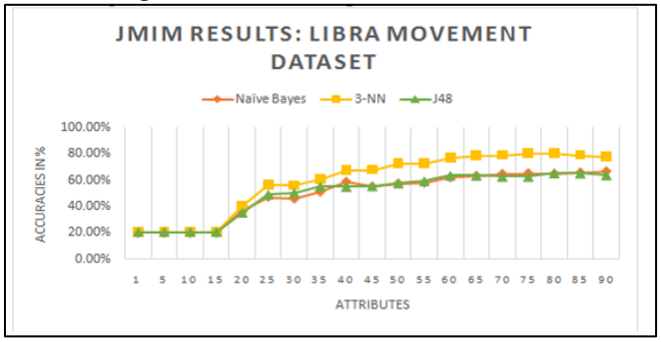

Fig. 4: JMIM results on Libra Movement Dataset

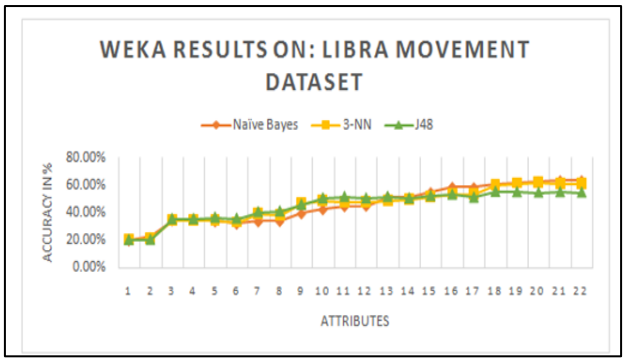

Fig. 5: WEKA results on Libra Movement Dataset

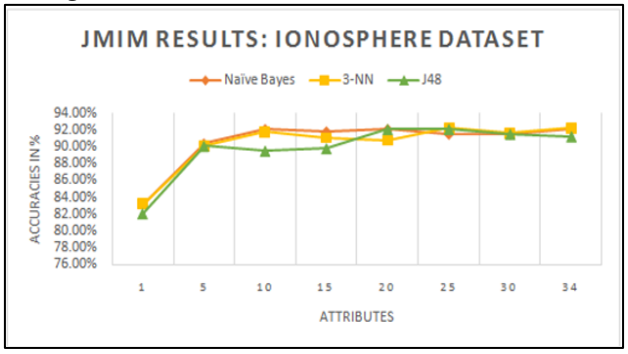

Fig. 6: JMIM results on Ionosphere Dataset

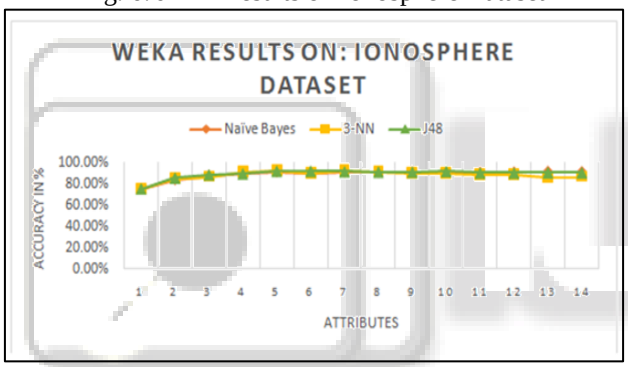

Fig. 7: WEKA results on Ionosphere Dataset

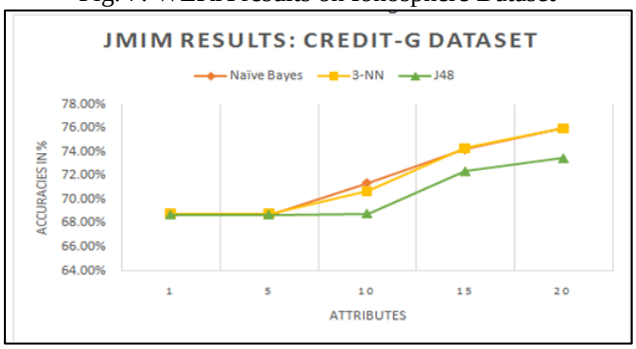

Fig. 8: JMIM results on Credit-G Dataset

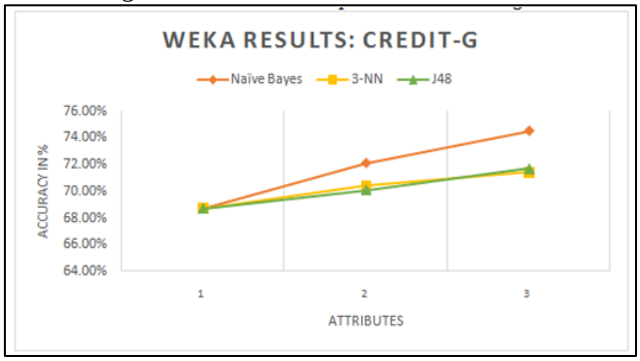

Fig. 9: WEKA results on Credit-G Dataset

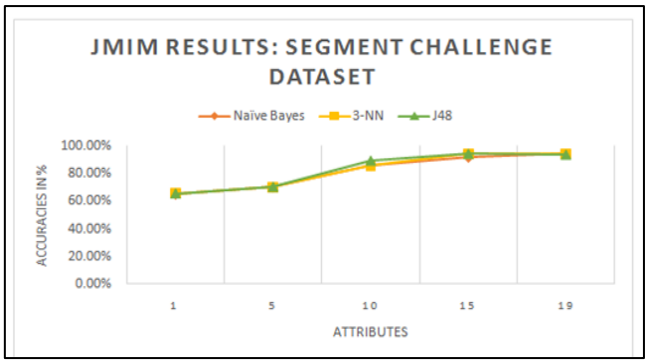

Fig. 10: JMIM results on Segment Challenge Dataset

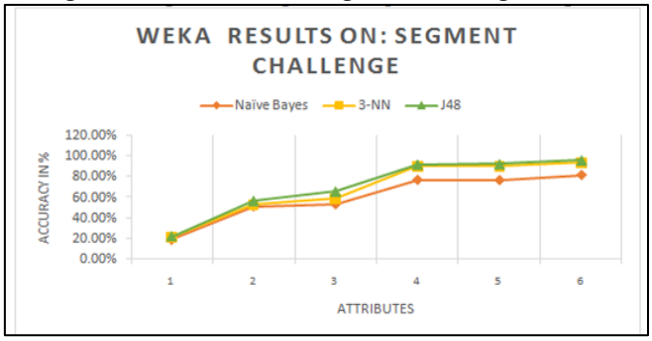

Fig. 11: WEKA results on Segment Challenge Dataset

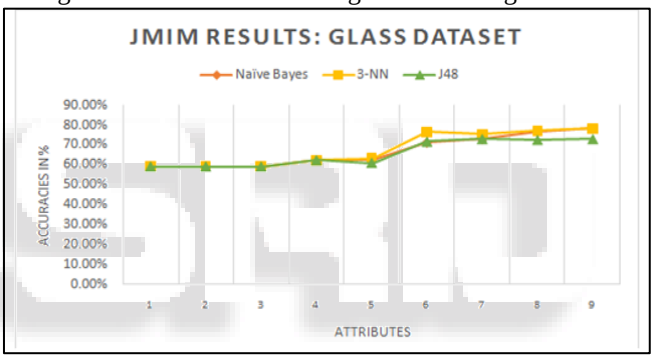

Fig. 12: JMIM results on Glass Dataset

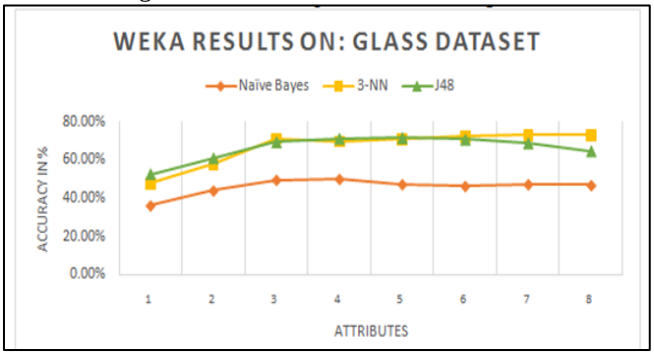

Fig. 13: WEKA results on Glass Dataset

Above shown graphs give the details of performance over our algorithm, JMIM and on WEKA – Attribute Selection. Below is the table that gives the comparative summary of the performance of JMIM and WEKA Attribute Selection.

Dataset	No. of Optimum Features			Accuracy Obtained		
	Naïve Bayes	3-NN	J48	Naïve Bayes	KNN	J48
Sonar	53	57	47	87.50%	88.46%	83.65%
Libra	90	77	82	66.67%	80.69%	66.39%
Ionosphere	9	28	17	92.31%	92.74%	92.02%
Credit G	19	18	20	77.00%	76.80%	73.50%
Segment Challenge	19	16	12	94.80%	94.83%	94.13%
Glass	9	9	7	78.27%	78.27%	72.90%

Table 5: JMIM results along with the optimum features

Dataset	No. of Optimum Features			Accuracy Obtained		
	Naïve Bayes	3-NN	J48	Naïve Bayes	3-NN	J48
Sonar	10	13	17	71.63%	90.38%	80.76%
Libra	21	20	18	63.88%	62.22%	55.55%
Ionosphere	13	7	5	91.73%	92.59%	92.02%
Credit G	3	3	3	74.50%	71.40%	71.70%
Segment Challenge	6	6	6	81.60%	93.73%	96.06%
Glass	4	7	5	50%	72.89%	71.49%

Table 6: WEKA results along with the optimum features

V. Observations

It is observed that for Sonar Dataset where JMIM gives 87.50% accuracy, WEKA gives accuracy of 71.63%. But this results of JMIM are obtained with 53 features whereas WEKA gives the accuracy of 71.63% with 10 features. This accuracy is obtained using Naïve Bayes classifier. For Libra Movement dataset, where WEKA gives accuracy of 62.22%, JMIM gives accuracy of 80.69%. Again the case is same as Sonar dataset. With Ionosphere dataset, the results approximately equal but, JMIM gives those results with high number of features. For Credit-G dataset, the results of JMIM are better, but are not very much greater than WEKA. For Segment Challenge dataset, WEKA gives results of 81.60% and JMIM gives the result of 94.80%. But the result of WEKA requires 6 features whereas the JMIM requires 19 features. Lastly, for the Glass dataset WEKA generates the results of 50% and JMIM gives results of 78.27%. WEKA gives these results with 4 features, and JMIM gives the accuracy over 9 features.

VI. Conclusion and Future Scope

Till now researchers have used mutual information in the process of feature selection for many algorithms. JMIM is the algorithm where this mutual information is being used with the 'maximum of minimum' approach. From the above summary we can say that JMIM gives better results in maximum experiments. We can also observe that in the most results 3-NN classifier gives highest results in terms of accuracy. But we cannot overlook the fact that the feature set required by JMIM is much greater as compared to WEKA tool to give these higher accuracy results. Hence, JMIM would be beneficial for its method of selection of features, since it assures the selection of most relevant features in classification tasks. In addition to the analysis of the public datasets used in this experiment, the algorithm could be used in many other applications where the relevance of the features for the classification task needs to be analysed. JMIM helps in analysis and detailed study of the feature selection process and hence would be beneficial if the higher size of feature sets is acceptable.

A comparative study of JMIM and Attribute Selection technique used in WEKA tool has been given. The time complexity is not measured in this research work. Hence, the time complexity can be considered in the future enhancement.

To reduce the time complexity, the system can be implemented in parallel environment. This will gradually generate the expected results in less time.

References

[1] Mohamed Bennasar, Yulia Hicks and Rossitza Setchi, "Feature selection using Joint Mutual Information Maximisation", Expert Systems with Application, Volume 42, Issue 22, 1 December 2015.
[2] Isabelle Guyon and Andr´e Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3 (2003) 1157-1182 2003.
[3] Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324
[4] Lal KN, Chapelle O, Weston J, Elisseeff A (2006) Embedded methods. In: Feature extraction, foundations and applications, studies in fuzziness and soft computing, vol 207, Springer, Berlin, chap 5, pp 167–185
[5] Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V (2000) Feature selection for svms. In: Advances in neural.
[6] Information processing systems 13, MIT Press, pp 668–674
[7] Duch W, Winiarski T, Biesiada J, Kachel A (2003) Feature selection and ranking filter. In: International conference on artificial neural networks (ICANN) and International conference on neural information processing (ICONIP), pp 251–254.
[8] Roberto Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 5, NO. 4, JULY 1994.
[9] Almuallim H, Dietterich TG (1991) Learning with many irrelevant features. In: Artificial intelligence, proceedings of the ninth national conference on, AAAI Press, pp 547–552
[10] Raudys S, Jain A (1991) Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 13(3):252–264
[11] Brown G, Pocock A, Zhao MJ, Luja´n M (2012) Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res 13:27–66
[12] Yang HH, Moody J (1999) Feature selection based on joint mutual information. In: Advances in intelligent data analysis, proceedings of international ICSC symposium, pp 22–25
[13] Webb AR (2002) Statistical Pattern Recognition. 2nd edn. Wiley, New Jersey
[14] Whitney A (1971) A direct method of nonparametric measurement selection. IEEE Trans Comput C-20(9):1100–1103
[15] Marill T, Green D (1963) On the effectiveness of receptors in recognition systems. IEEE Trans Inform Theory 9(1):11–17
[16] M. Dash, H. Liu,"Feature Selection for Classification,"ELSEVlER -Intelligent Data Analysis 1 (1997) 131-156.

