Application of Feature Weighting for the Intensification of Data Classification

J. Arunadevi, K. Ganeshamoorthi, R. Rampriya

Abstract*: Classification is the supervised learning technique which is applied in many of the real time applications. In this study we have considered three classifiers which are widely used and then the intensification of the classifiers are considered. Among various methods to improve the performance of the classifiers, this research concentrate on the feature weighting techniques applied for the classifiers. This analysis is done based on the results obtained from the Rapidminer tool. Here we have deployed four feature weighting techniques for the intensification of the three classifiers. It is tested with three dataset. The experimental environment and the results are discussed in detail.

***Keywords* : Feature weighting, Decision Tree, KNN, Naïve bayes**

I. INTRODUCTION

The proper usage of data preprocessing reduces the complexity and increase the performance of any machine learning algorithm. Here we are going to concentrate on the feature weighting technique to intensify the classification process. The feature weighting could be done by several methods. In this research we have concentrated on four different types of feature weighting methods. It has been tested with three classifiers. The research experiment is to be carried out for the three dataset.

II. FEATURE WEIGHTING

To avoid the problem of high dimensionality we can employ many methods in machine learning. Feature weighting is one of the processes which is used to solve the high dimensionality. Over fitting occur when small data sets are contained with large dimension of data [1-3]. In feature weighting features are allocated different weights to replicate their relevance to the output. Mostly irrelevant or redundant features are assigned a very low weight value. Feature weighting could be used in supervised learning for the development of methods for computing the capability of features to distinguish instances from diverse classes [4-5]. It computes class dependent feature weight vectors.

III. DATA CLASSIFICATION

Classification is the supervised machine learning algorithm, which is used to group the data items based on the label provided for the instance. This could be carried out in two phases. The first phase is the training phase and the second phase is the testing phase. In the training phase the algorithm is trained by the data instances which is grouped based on the label. In the testing phase the algorithm is tested to identify the data instances on the label.

IV. INTENSIFICATION OF DATA CLASSIFICATION

Data classification is the supervised learning task which is used to group the given data based on the knowledge acquired from training of the data. But the classification task suffers by not reaching the accurate classification of the data items [6-9]. So it is important to find the ways to improve the accuracy and other parameters related to the classification task like kappa, precision and recall. This paper suggests a way to intensify the classification task by applying feature weighting technique and intensify the classification

V. PROBLEM FORMULATION

The classification of the data is an important task in many real world applications. We have discussed three classifiers in the study and there is a scope of research in this area about consideration of the features in the dataset towards the classification task. In this work we would like to apply some feature weighting methods to the dataset and then classify the resultant data.

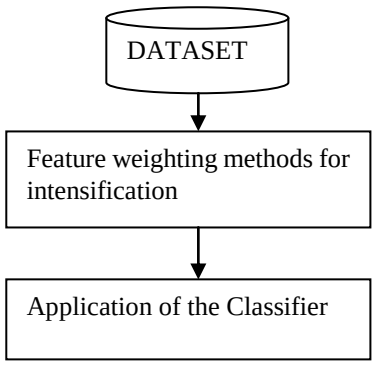


Fig 1 : Workflow of the proposed work

VI. DATASET USED

There are three dataset has been used for this experiment which are available in Rapidminer.

Revised Manuscript Received on December 16, 2019.
***** Correspondence Author
Dr.J.Arunadevi*, Assistant Professor, Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, Affilliated to Alagappa Univesity, Tamilnadu, India, arunasethupathy@gmail.com
K.Ganeshamoorthi, Ph.D Scholar (Full-Time), Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, Affilliated to Alagappa Univesity, Tamilnadu, India
R.RAMPRIYA, M.Phil Scholar (Full-Time), Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, Affilliated to Alagappa Univesity, Tamilnadu, India

Table 1: Dataset description

Name of the Dataset	No. of attributes	No. of examples
Deals	4	1000
Golf	5	14
Sonar	61	208

VII. EXPERIMENTAL DETAILS

Three dataset with three classifier algorithms for the four feature weighting algorithms to implement is presented here. The results are obtained in the following combinations.

Table 2: Details of the experiments conducted

Dataset used	Feature Weighting method used	Classification algorithm used
Deals, Golf & Sonar	Information Gain	Decision Tree
		K Nearest Neighbors
		Naive Bayes
Deals, Golf & Sonar	Principal Component Analysis	Decision Tree
		K Nearest Neighbors
		Naive Bayes
Deals, Golf & Sonar	Genetic Algorithm	Decision Tree
		K Nearest Neighbors
		Naive Bayes
Deals, Golf & Sonar	Particle Swarm Optimization	Decision Tree
		K Nearest Neighbors
		Naive Bayes

VIII. PERFORMANCE METRICS USED

The performance metrics used for the experiment is given below

- Accuracy - Accuracy is how close a measured value is to the true value. It expresses the correctness of a measurement and determined by absolute and comparative way.
- Error - Relative number of misclassified examples or in other words percentage of incorrect predictions.
- Kappa - The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance).
- Weighted mean recall - The weighted mean of all per class recall measurements. It is calculated through class recalls for individual classes.
- Weighted mean precision - The weighted mean of all per class precision measurements. It is calculated through class precisions for individual classes

IX. RESULTS AND DISCUSSIONS

The following tables gives the details of the consolidated results obtained by the experiments conducted, which is discussed above.

Tables 3, 4 and 5 shows the results obtained for the classification performance metric without the application of feature weighting techniques

Tables 6, 7, 8 and 9 gives the results for the application of the four feature weighting techniques to the Deals dataset.

Tables 10, 11, 12 and 13 gives the results for the application of the four feature weighting techniques to the Golf dataset.

Tables 14, 15, 16 and 17 gives the results for the application of the four feature weighting techniques to the Sonar dataset.

Table 3 Results obtained when classification without feature weighting applied on Deals dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	99.6	0.4	0.992	99.62	99.59
KNN	97.3	27	0.946	97.4	97.33
Naïve	92.6	7.4	0.852	92.64	92.64

Table 4 Results obtained when classification without feature weighting applied on Golf dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	55	45	-0.024	35	27.5
KNN	40	60	-0.143	27.5	22.5
Naïve	55	45	0.364	42.5	45

Table 5 Results obtained when classification without feature weighting applied on Sonar dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	62.12	37.88	0.231	61.45	62.61
KNN	66.9	33.1	0.348	67.76	69.79
Naïve	82.14	17.86	0.639	81.68	84.21

Table 6 Results obtained when Information gain feature weighting applied on Deals dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	99.5	0.5	0.99	99.52	99.5

KNN	98	2	0.96	98.09	97.96
Naïve	93	7	0.858	92.91	93.01

Table 7 Results obtained when PCA weighting applied on Deals dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	99.7	0.3	0.994	99.69	99.7
KNN	77.2	2.8	0.541	77.19	77.2
Naïve	88.1	11.9	0.76	88.01	88.15

Table 8 Results obtained when GA weighting applied on Deals dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	99.6	0.4	0.992	99.61	99.6
KNN	99.6	0.4	0.992	99.62	99.59
Naïve	92.7	7.3	0.854	92.74	92.87

Table 9 Results obtained when PSO weighting applied on Deals dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	99.9	0.1	0.998	99.91	99.9
KNN	100	0	1	100	100
Naïve	93.2	6.8	0.864	93.21	93.19

Table 10 Results obtained when Information Gain weighting applied on Golf dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	50	50	0.039	32.5	32.5
KNN	70	30	0.378	42.5	42.5
Naïve	55	45	-0.024	32.5	30

Table 11 Results obtained when PCA weighting applied on Golf dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	55	45	-0.256	35	27.5
KNN	40	60	-0.235	25	20
Naïve	50	50	-0.14	32.5	27.5

Table 12 Results obtained when GA weighting applied on Golf dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	70	30	0.186	42.5	37.5
KNN	45	55	-0.366	27.5	25
Naïve	70	30	0.186	45	40

Table 13 Results obtained when PSO weighting applied on Golf dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	75	25	0.378	47.5	45

KNN	90	10	0.659	55	50
Naïve	70	30	0.186	42.5	42.5

Table 14 Results obtained when Information gain weighting applied on Sonar dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	74.02	25.98	0.465	73.59	74.18
KNN	74.57	25.43	0.483	74.13	75.12
Naïve	74.52	25.48	0.477	74.74	74.81

Table 15 Results obtained when PCA weighting applied on Sonar dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	54.93	45.07	0.041	51.98	47.26
KNN	75.95	24.05	0.514	75.94	77.17
Naïve	63.55	36.45	0.255	62.69	63.87

Table 16 Results obtained when GA weighting applied on Sonar dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	67.33	32.67	0.338	66.77	68.14
KNN	87.93	12.07	0.755	87.59	89.19
Naïve	70.74	29.26	0.422	71.47	74

Table 17 Results obtained when PSO weighting applied on Sonar dataset

Classifier	Accuracy	Error	Kappa	WM recall	WM precision
DT	72.14	27.86	0.435	71.58	73.15
KNN	92.74	7.26	0.854	92.61	93.06
Naïve	72.12	27.88	0.446	72.54	74.99

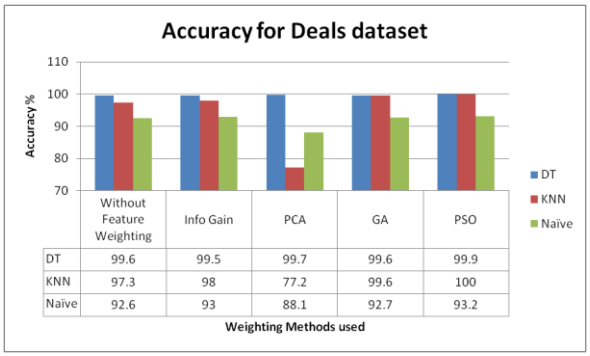


Fig 2: Comparison based on Accuracy for Deals dataset

Fig 2 displays the comparison of various feature weighting methods when it is applied to the deals dataset. In terms of accuracy The PSO outperforms other methods when it is combined with the KNN classifier.

Fig 3 displays the comparison of various feature weighting methods when it is applied to the golf dataset. In terms of accuracy The PSO outperforms other methods when it is combined with the KNN classifier.

Fig 4 displays the comparison of various feature weighting methods when it is applied to the sonar dataset. In terms of accuracy The PSO outperforms other methods when it is combined with the KNN classifier.

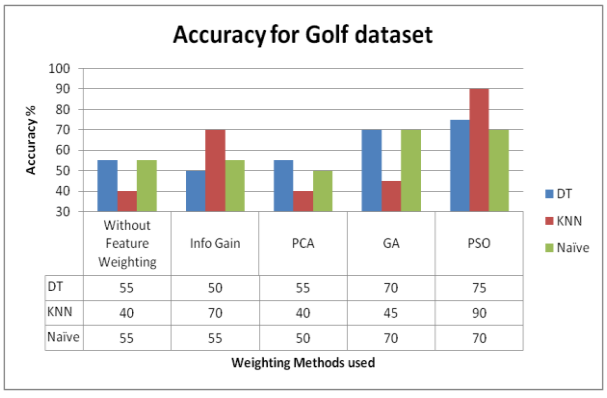


Fig 3: Comparison based on Accuracy for Golf dataset

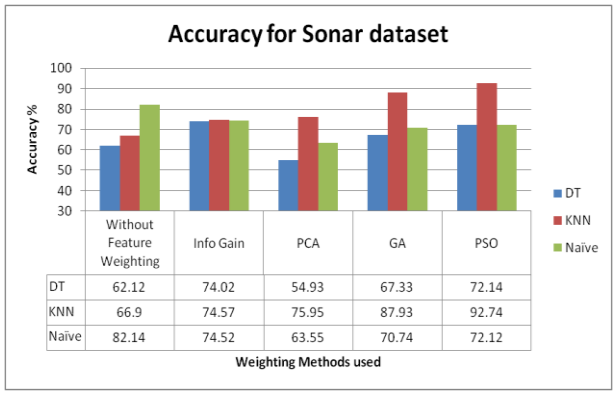


Fig 4: Comparison based on Accuracy for Sonar dataset

Fig 5 displays the comparison of various feature weighting methods when it is applied to the Deals dataset. In terms of Kappa statistics the PSO outperforms other methods when it is combined with the KNN classifier.

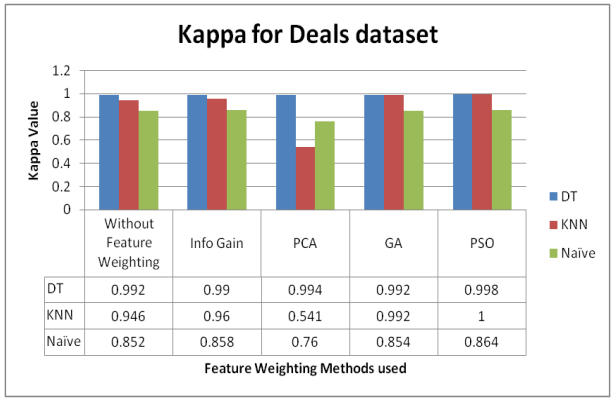
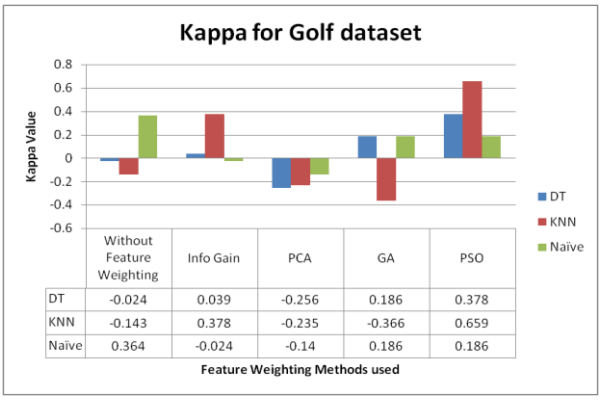


Fig 5 : Comparison based on Kappa for Deals dataset

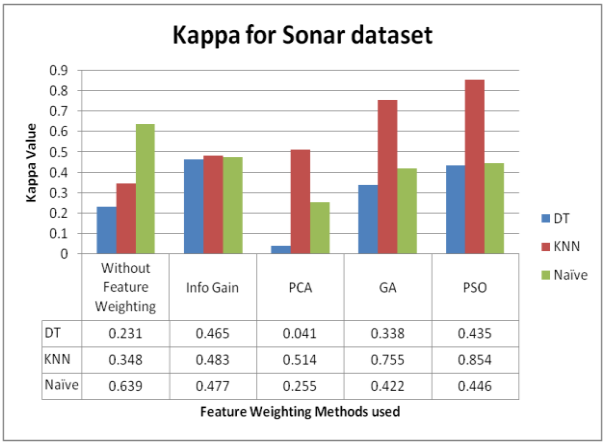
Fig 6 displays the comparison of various feature weighting methods when it is applied to the Golf dataset. In terms of Kappa statistics the PSO outperforms other methods when it is combined with the KNN classifier.

Fig 7 displays the comparison of various feature weighting methods when it is applied to the Sonar dataset. In terms of Kappa statistics the PSO outperforms other methods when it is combined with the KNN classifier.



	Without Feature Weighting	Info Gain	PCA	GA	PSO
DT	-0.024	0.039	-0.256	0.186	0.378
KNN	-0.143	0.378	-0.235	-0.366	0.659
Naïve	0.364	-0.024	-0.14	0.186	0.186

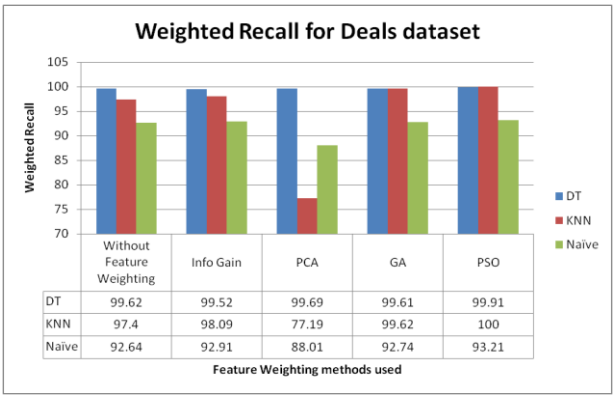
Fig 6: Comparison based on Kappa for Golf dataset



	Without Feature Weighting	Info Gain	PCA	GA	PSO
DT	0.231	0.465	0.041	0.338	0.435
KNN	0.348	0.483	0.514	0.755	0.854
Naïve	0.639	0.477	0.255	0.422	0.446

Fig 7: Comparison based on Kappa for Sonar dataset

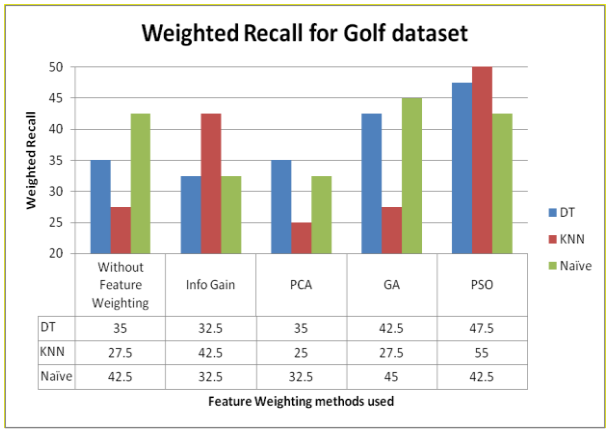
Fig 8 displays the comparison of various feature weighting methods when it is applied to the Deals dataset. In terms of Weighted Recall the PSO outperforms other methods when it is combined with the KNN classifier.



	Without Feature Weighting	Info Gain	PCA	GA	PSO
DT	99.62	99.52	99.69	99.61	99.91
KNN	97.4	98.09	77.19	99.62	100
Naïve	92.64	92.91	88.01	92.74	93.21

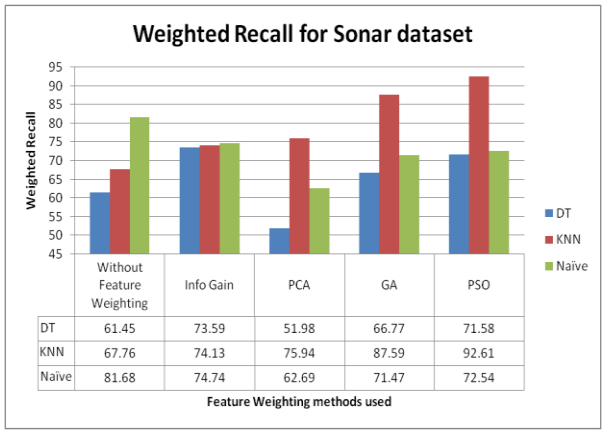
Fig 8 : Comparison based on Weighted Recall for Deals dataset

Fig 9 displays the comparison of various feature weighting methods when it is applied to the Golf dataset. In terms of Weighted Recall the PSO outperforms other methods when it is combined with the KNN classifier.
Fig 10 displays the comparison of various feature weighting methods when it is applied to the sonar dataset. In terms of Weighted Recall the PSO outperforms other methods when it is combined with the KNN classifier.



	Without Feature Weighting	Info Gain	PCA	GA	PSO
DT	35	32.5	35	42.5	47.5
KNN	27.5	42.5	25	27.5	55
Naïve	42.5	32.5	32.5	45	42.5

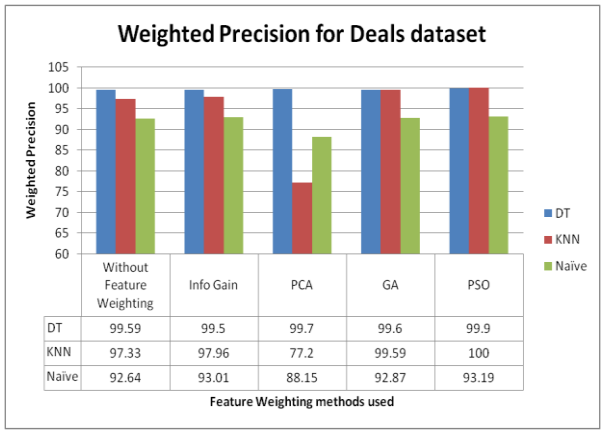
Fig 9: Comparison based on Weighted Recall for Golf dataset



	Without Feature Weighting	Info Gain	PCA	GA	PSO
DT	61.45	73.59	51.98	66.77	71.58
KNN	67.76	74.13	75.94	87.59	92.61
Naïve	81.68	74.74	62.69	71.47	72.54

Fig 10: Comparison based on Weighted Recall for Sonar dataset

Fig 11 displays the comparison of various feature weighting methods when it is applied to the Deals dataset. In terms of Weighted Precision the PSO outperforms other methods when it is combined with the KNN classifier.
Fig 12 displays the comparison of various feature weighting methods when it is applied to the Golf dataset. In terms of Weighted Precision the PSO outperforms other methods when it is combined with the KNN classifier.
Fig 13 displays the comparison of various feature weighting methods when it is applied to the Sonar dataset. In terms of Weighted Precision the PSO outperforms other methods when it is combined with the KNN classifier.



	Without Feature Weighting	Info Gain	PCA	GA	PSO
DT	99.59	99.5	99.7	99.6	99.9
KNN	97.33	97.96	77.2	99.59	100
Naïve	92.64	93.01	88.15	92.87	93.19

Fig 11: Comparison based on Weighted Precision for Deals dataset

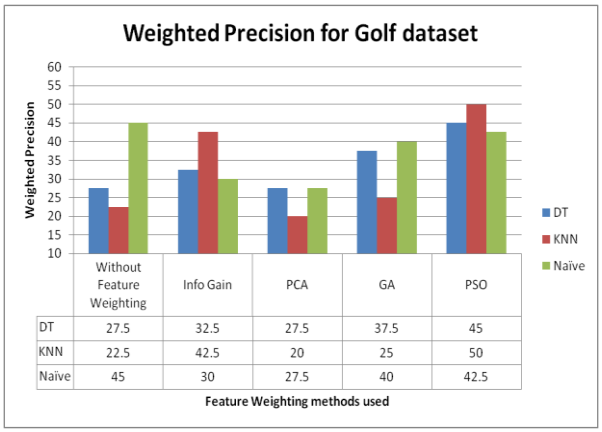


Fig 12 Comparison based on Weighted Precision for Golf dataset

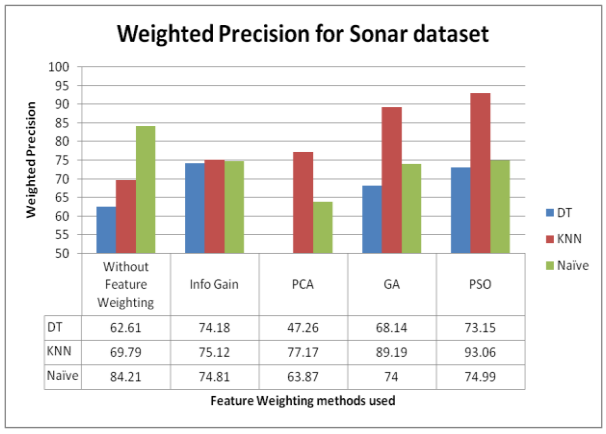


Fig 13 Comparison based on Weighted Precision for Sonar dataset

The results obtained are discussed and the best results based on the various performance measures obtained from the classifiers are discussed and compared with the feature weighted facilitated classifiers. The results are tabulated below. The analysis of the results shows that the application of feature weighting for the classification algorithms improves the classification performance measures.

Table 18 Comparison of the feature facilitated classifier experiment with the bench marks results

Dataset	Classifier	Accuracy	Error	Kappa	WM Recall	WM Precision
Deals	DT	99.6	0.4	0.992	99.62	99.59
	PSO+KNN	**100**	**0**	**1**	**100**	**100**
Golf	Naive	55	45	0.364	42.5	45
	PSO + KNN	**90**	**10**	**0.659**	**55**	**50**
Sonar	Naïve	82.14	17.86	0.639	81.68	84.21
	PSO + KNN	**92.74**	**7.26**	**0.854**	**92.61**	**93.06**

X. CONCLUSION AND FUTURE ENHANCEMENT

This work discusses about the various feature weighting methods for the intensification of the classifier. This report gives the information on the types and the usage of it. This research concentrates on the analysis of the classifier and its intensification using the feature weighting methods it to the standard dataset in the rapid miner tool. Based on the experimental results KNN classifier with the PSO based feature weighting produces the better classification for the dataset considered. The future work could be concentrated on considering more methods for the intensification of the classifier and to be tested on real time dataset.

REFERENCES

1. Sujoy Paul ,SwagatamDas, Simultaneous feature selection and weighting – An evolutionary multi-objective optimization approach, Pattern Recognition Letters 65 (2015) 51–59
2. https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/feature-selection-data-mining
3. A. Ko lcz and W.-T. Yih. Raising the baseline for high-precision text classifiers. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , pages 400–409. ACM, 2007.
4. Elena Marchiori, Class dependent feature weighting and K-nearest neighbor classification. http://www.cs.ru.nl/~elenam/prib2013_asymmetric_submission.pdf
5. 5. Janez Brank, Natasa Milic-Frayling, A Framework for Characterizing Feature Weighting and Selection Methods in Text Classification, Tech report, Microsoft 2005
6. A. Ko lcz and W.-T. Yih. Raising the baseline for high-precision text classifiers. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , pages 400–409. ACM, 2007.
7. Elena Marchiori, Class dependent feature weighting and K-nearest neighbor classification, http://www.cs.ru.nl/~elenam/prib2013_asymmetric_submission.pdf
8. Songtao Shang, Minyong Shi, Wenqian Shang, and Zhiguo Hong, "Improved Feature Weight Algorithm and Its Application to Text Classification," Mathematical Problems in Engineering, vol. 2016.
9. Christos Boutsidis, Michael W. Mahoney, Petros Drineas, Unsupervised Feature Selection for Principal Components Analysis, KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

AUTHORS PROFILE

Dr.J. Arunadevi, working as the Assistant Professor is interested in the area of machine learning and soft computing. She is activly involved in the research which is concentrated on the use of soft computing techniques. She has several publications in the International Journals and Conferences.

K.Ganeshamoorthi is the full time PhD research scholar working in the area of data reduction and interested in feature engineering

R.Rampriya is the MPhil scholar, with the research area feature weighting for the data preprocessing works.