Feature Selection in Meta Learning Framework

Samar Shilbayeh
Department of Computer science and Engineering
University of Salford.
Manchester, UK.
s.a.shilbayeh@edu.salford.ac.uk

Sunil Vadera
Department of Computer science and Engineering
University of Salford
Manchester, UK
S.Vadera@salford.ac.uk

***Abstract*—Feature selection is a key step in data mining. Unfortunately, there is no single feature selection method that is always the best and the data miner usually has to experiment with different methods using a trial and error approach, which can be time consuming and costly especially with very large datasets. Hence, this research aims to develop a meta learning framework that is able to learn about which feature selection methods work best for a given data set. The framework involves obtaining the characteristics of the data and then running alternative feature selection methods to obtain their performance. The characteristics, methods used and their performance provide the examples which are used by a learner to induce the meta knowledge which can then be applied to predict future performance on unseen data sets.**

This framework is implemented in the Weka system and experiments with 26 data sets show good results.

Keywords—Meta learning; feature selection; supervised classification; algorithim selection

I. INTRODUCTION

A central problem in data mining is to identify which features of the data are most useful for obtaining good results [1,2]. Hence, many methods have been developed to improve the feature selection process, such as wrapper methods [3,4, 5], filter methods[6,7], and methods that use fuzzy rough sets[8]. Unfortunately, there is no dominating feature selection method that works best in all cases [9].

One direction of research is to continue to seek the ultimate feature selection method that always works well. Another approach, taken by this research, is to accept that one method does not fit all requirements, but instead aim to identify which method works best for a given data set. However, this is not easy, since details of which algorithm works best under different circumstances is not known. Thus, we have a meta learning problem, namely:

Can we automatically learn which feature selection algorithm works best for different circumstances?

This paper aims to answer this question by developing a new meta-learning framework that aims to learn from the experience of applying different feature selection methods on data sets with different characteristics.

The paper is organized as follows: Section 2 presents the related work on feature selection, Section 3 presents framework, Section 4 the results of the empirical evaluation, and Section 5 concludes the paper.

II. BACKGROUND AND RELATED WORK

This section summarizes some of the key research on feature selection and the reader is referred to several excellent accounts of feature selection methods, such as in [10, 11] as well as surveys [12, 13].

A general definition of feature selection is: the process of selecting a subset of features that maximizes the predictive power of a learner. However, feature selection has been covered by researchers from different angles:

- Find the subset of features that optimizes the evaluation functions including increasing a classifier's performance or decreasing the computational cost without reducing its performance.
- Find the subset of features that has a direct relation to the target label [2].

Many feature selection methods and search strategies have been proposed in the literature (e.g., [1, 2, and 11]). The main two kinds of feature selection methods identified include: (i) a filter approach which deals with the data characteristics independently from the learning classifier and (ii) a wrapper approach which uses the classifier itself to identify the subset of features that are more effective in the learning process.

Two of the earliest algorithms that adopt filtering include FOCUS[1] and RELIEF[2]. The FOCUS algorithm starts with an empty set and uses a breadth-first search strategy to find the optimal solution, whereas the RELIEF algorithm assigns a weight to each feature then finds the optimal feature that exceeds a user threshold. Both of these filtering methods adopt ID3 to induce a decision tree by using the selected attributes after the feature selection process. FOCUS searches for a minimal set of features, while RELIEF searches for all relevant features. Those two algorithms assumed Boolean attributes, whereas Molinaet al. [12] reports extensions to their methods that handle non-Boolean attributes and multiple classes. Liu and Rudy [13] develop the LVF algorithm for filtering based on probabilities and use a consistency measure that is different from that of FOCUS, which improves the ability for finding optimal subsets even with noisy data.

Other recent works rely on the wrapper approach rather than filter approach.

The main argument for using the wrapper method is that the inducer itself is used in the feature selection process and this will reduce induction bias. This is in contrast to the filter approach where two different isolated strategies are adopted: one for feature selection and one for induction.

The first pioneering work that uses the wrapper approach in feature selection was by George et al. [14] in which a decision tree is used to find good feature subset. The idea is that a learner is applied on a chosen subset of features and its accuracy obtained. Features can then be added and removed depending on their effect on the accuracy. Several authors have built upon this basic idea and used different approaches for searching for the subset of features, so for example,

Skalak et al. [15] use hill-climbing search, Pudil et al. [11] use a floating search strategy, and Yang and Vasant [16] use a genetic algorithm for feature subset selection.

There have been number of comparative evaluations for different feature selection approaches in different application such as the study by Vasantha et al. [17] on mammogram images, and by Liu and Schumann [18] who used feature selection approaches and techniques to increase the performance of a credit scoring model.

As the above suggests there are many different feature selection methods and no doubt the list will continue to grow in the future. In our work the importance of feature selection exists in exploring which kind of feature selection methods most suit a specific types of data. That is, we are interested in learning the relationship(s) between feature selection methods, data set characteristics and performance as expressed in accuracy and the misclassification cost.

The next section proposes a framework for meta learning that we are developing and which includes learning about feature selection methods.

III. A New Meta Learning Framework

Given the above motivation, this research aims to build upon existing work and develop a meta learning framework. The main idea of the framework is to use meta features to characterize the data and learn the performance of different algorithms. This learned knowledge is then used by a planner to develop a suitable plan for a given situation; it is worth to point out here that this paper covers part of this framework which is developing a feature selection meta knowledge that selects the most suitable feature selection plan for a given data set. Fig.1 presents the proposed framework which has the following components

1) Meta feature: Applying different data characterization techniques on a given example data sets to understand the data behavior and nature, this includes different data set characterization approaches such as simple, statistical , mathematical and landmarking characterization [19].

2) Feature selection: applying different feature selection approaches and search strategies to build knowledge on which feature selection plan will suit a specific data set.

3) Cost sensitive learning: applying different cost sensitive and insensitive approaches for the aim of building a model that predicts the classifier performance and cost for a given data set taking into consideration the classifier misclassification cost.

4) Performance evaluation: Applying different algorithms on a given example data set and evaluate the result is the core of this stage, different evaluation criterion are used such as accuracy, and misclassification cost

5) Meta learning process: The main goal of this phase is to learn about the data mining process which includes the result of all previous phases: data sets of examples with its characteristics (meta features), different feature selection strategies, and all sets of algorithms with their performance after applying cost sensitive and insensitive learning all fed to the meta learner for the aim of developing a meta knowledge that guide the data mining process.

6) Planning: Applying different machine learning algorithms successfully in data mining often involves acquiring the data, pre-processing, and choosing the best algorithm; thus, ideally one needs to plan how a particular problem will be tackled.

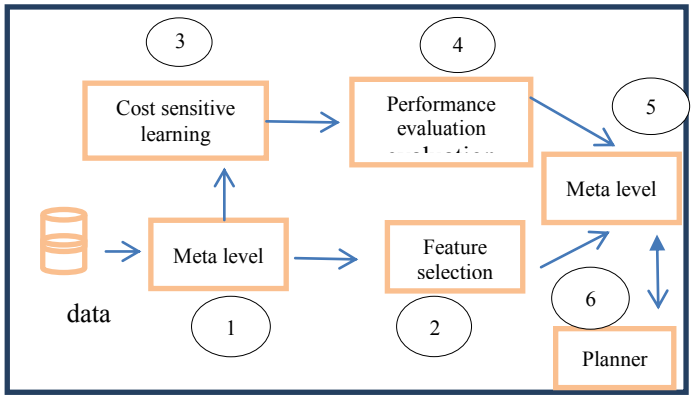

Fig. 1. New meta learning framework

In this paper part of this framework will be implemented and tested which includes the following:

1) Meta features: obtain the meta features for each data set

2) Feature selection: apply the different feature selection methods to generate the table of examples

3) Performance evaluation: includes applying different classifiers to evaluate classifiers performance and cost

4) Meta knowledge: apply J4.8 to generate the decision tree.

Fig.1.a presents the feature selection meta knowledge development component that imp

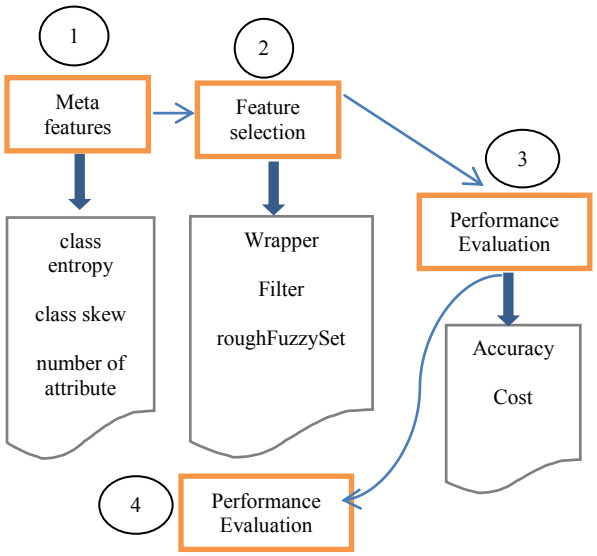

Fig. 1. a Feature selection meta knowledge development

IV. EMPIRICAL EVALUATIONS

The empirical evaluations were carried out in two stages. The first experiments, presented in A involved simply assessing which methods worked well and also if there was in fact a best method. The second experiments, presented in B involved learned and applied the meta knowledge for recommending feature selection methods.

The experiments in this section are carried out using 26 data sets. All the data sets are taken from the UCI Machine Learning repository, the evaluations were carried out using 10 cross validations.

A. Feature selection evaluation

To evaluate the effect of using feature selection methods on classifier performance, a comparison is made between using and not using a feature selection method. Four very different methods are considered and the results presented below.

TABLE I summarizes the improvements in classifiers accuracy after using a wrapper method, known as the WrappersubsetEval method in Weka [20] , that uses a greedy search method for adding/removing attributes. The results are presented in three columns for each base learning method: a column with the accuracy prior to use of feature selection, a column after utilizing the wrapper method with feature subset selection (columns with a FS suffix), and a column showing the improvement (columns labelled IM). Results are presented for a decision tree learner known as J4.8, Naive Bayes (NB), and neural networks (NW).

TABLE II summarizes the improvements in classifiers performance using a filtering method that uses worth of an attribute by using an information theoretic measure that is used in decision tree learning algorithms known as the gain ratio with respect to the class. .

TABLE III summarizes the improvements in classifier performance using a method, known as cfsSubEval in Weka, that evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them [21]

TABLE VI summarizes the improvements in classifiers accuracy after using a novel method that uses rough set theory for assessment of the quality of a subset of features together with Particle Swarm Optimization for subset optimization [22,23].

TABLE I. CHANGES IN CLASSIFIER PERFORMANCE AFTER USING WRAPPERSUBSETEVAL WITH GREEDYSEARCH

Data set	J48	J48 FS	IM	NB	NB FS	IM	NW	NW FS	IM
Contact Lenses	83.3	83.3	0	70.8	70.8	0	70.8	75	4.1
Credit-Card g	70.5	74.6	4.1	75.4	74.4	-1	71.6	73.5	1.9
Diabetes	73.8	75	1.2	76.3	77.5	1.2	75.3	77.2	1.8
Glass	66.6	68.6	2.0	48.5	55.1	6.5	67.7	66.8	-0.9
Ionosphere	91.4	90.5	-0.8	82.6	92.0	9.4	91.1	93.4	2.2
Labor	73.6	80.7	7.0	89.4	91.2	1.7	85.9	89.4	3.5
Weather	64.2	.7	7.1	64.2	64.2	0	78.5	78.5	0
Soybean	91.5	92.9	1.4	92.9	93.8	0.9	93.4	93.4	0
Vote	96.3	95.6	-0.7	90.1	96.3	6.2	94.1	97.4	3.3
Cancer	75.5	75.8	0.3	71.6	75.1	3.5	64.6	75.8	11.2
Average of IM			2.17			2.84			2.73

TABLE II. CHANGES IN CLASSIFIER PERFORMANCE AFTER USING GAINRATIOEVAL WITH RANKER

Data set	J48	J48 FS	IM	NB	NB FS	IM	NW	NW FS	IM
Contact Lenses	83.3	83.3	0	70.8	83.3	12.4	70.8	83.3	12.4
Credit-Card g	70.5	72.8	2.3	75.4	75.5	0.1	71.6	71.4	-0.2
Diabetes	73.8	74.0	0.2	76.3	75.1	-0.01	75.3	75.3	0
Glass	66.6	68.3	1.6	48.5	50	1.4	67.7	66.8	-0.9
Ionosphere	91.4	90.3	-1.1	82.3	87.1	4.8	91.1	91.1	0
Labor	73.6	77.1	3.5	89.4	89.4	0	85.9	82.5	-3.4
Weather	64.2	71.4	7.1	64.2	57.2	-0.07	78.5	78.5	0
Cancer	75.5	75.1	-0.03	71.6	72.3	12.4	64.6	67.3	12.4
Average of IM			0.6			2.3			2.02

TABLE III. CHANGES IN CLASSIFIER PERFORMANCE AFTER USING CFSSUBEVAL WITH BESTFIRST

Data set	J48	J48 FS	IM	NB	NB FS	IM	NW	NW FS	IM
Contact Lenses	83.3	70.8	-12.5	70.8	70.8	0	70.8	66.6	-4.2
Credit-Card g	70.5	70.5	0	75.4	74.4	-1	71.6	73	1.4
Diabetes	73.8	74.8	1.1	76.3	77.5	1.2	75.3	75.5	0.1
Glass	66.6	68.9	2.3	48.5	47.6	-0.9	67.7	65.8	-1.9
Ionosphere	91.4	90.5	-0.9	82.6	92.0	9.3	91.1	93.4	2.2
Labor	73.6	77.1	3.5	89.4	91.2	1.73	85.9	85.9	0
Weather	64.2	42.8	-21.4	64.2	57.1	-7.1	78.5	71.4	-7.1
Soybean	91.5	90.1	-1.4	92.9	92.2	-0.7	93.4	93.8	0.4
Vote	96.3	96	-0.3	90.1	96	5.9	94.1	95.8	1.7
Cancer	75.5	73.0	-2.5	71.6	72.3	0.7	64.6	71.6	7.07
Average of IM			-3.2			0.9			-.03

TABLE IV. CHANGES IN CLASSIFIER PERFORMANCE AFTER USING FUZZYSUBSETEVAL WITH PSOSEARCH

Data set	J48	J48 FS	IM	Naive Bayes	NB FS	IM	Neural Network	Neural FS	IM
Contact Lenses	83.3	83.3	0	70.8	70.8	0	70.8	70.8	0
Credit-Card g	70.5	71.2	0.7	75.4	73.6	-1.8	71.6	71.5	-0.1
Diabetes	73.8	73.8	0	76.3	76.3	0	75.3	75.3	0
Glass	66.6	67.7	1.1	48.5	42.5	-6	67.7	66.3	-1.4
Ionosphere	91.4	88.0	-3.3	82.6	90.0	7.4	91.1	90.5	-0.6
Labor	73.6	80.7	7.1	89.4	91.2	1.8	85.9	87.7	1.8
Weather	64.2	71.4	7.2	64.2	57.1	-7.0	78.5	78.5	0
Soybean	91.5	85.2	-6.3	92.9	97.5	4.6	93.4	85.2	-8.2
Vote	96.3	96.3	0	90.1	92.8	2.7	94.7	95.4	0.7
Cancer	75.5	75.5	0	71.6	73.0	1.4	64.6	68.2	3.6
Average of IM			6.43			3.06			-4.2

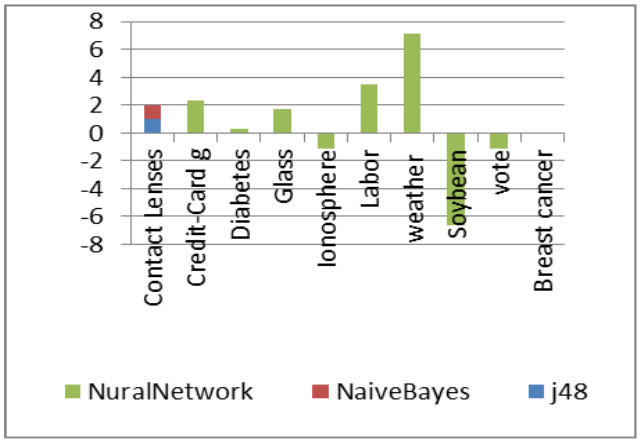

Fig. 2. Wrappe subsetEval with Greedysearch

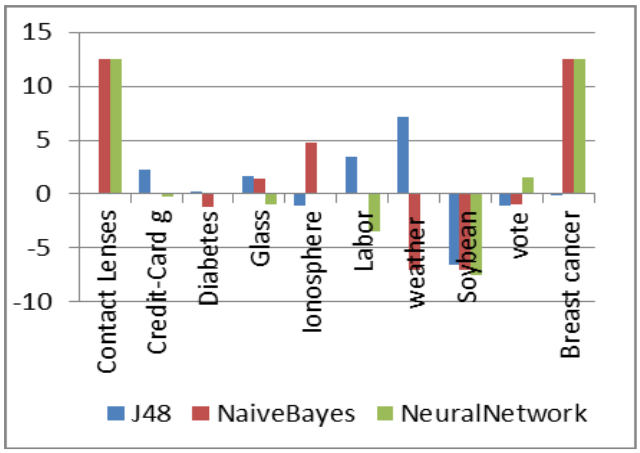

Fig. 3. GainRatioAttributeEval with Ranker

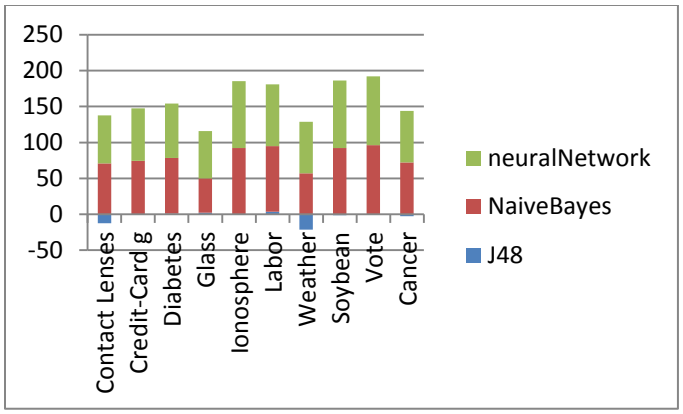

Fig. 4. cfsSubEval with BestFirst

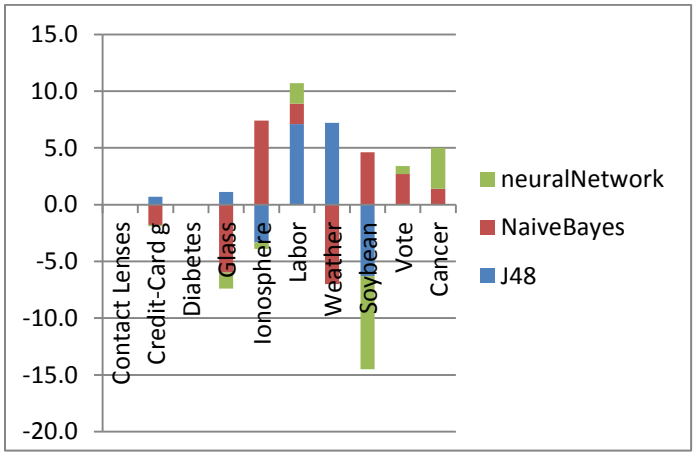

Fig. 5. FuzzyRoughSubSetEval with PSOSearch

B. Conclusion

Applying different attribute selection approaches with different search strategies on datasets and comparing the results of classifier performance between using and not using feature selection, the results show that in general there is improvement in the learning algorithm performance when using feature selection, and shows that there is no best feature selection approach for all learning tasks. For example, while using the wrapperSubsetEval approach with bestFirst search shows an improvement on J48, over NaiveBayes, and neuralNetworks on a specific data set, using cfsSubsetEval with best first strategy shows a decrease of performance in most of the tested data. Figs. 2, 3, 4 and 5 show the improvement and the decrements in different classifiers accuracy after using different feature selections approaches with different search strategies.

What is required for a good data mining plan is to understand the data nature, and to know exactly what is best pre-processing strategy that works in a specific data set.

Thus the aim of our next experiment is to link between data set characteristics, feature selection approaches and search different techniques with different classifiers performance to know which feature selection method is best for specific task.

V. META LEARNING STAGE

In this stage, 26 data sets from UCI [24] are used to develop a meta knowledge that guides the process of feature selection. For this aim, different data set characteristics are used for the purpose of understanding the nature of the data and to know what makes a specific attribute selection approach work well on a specific data set. For each data set the following characteristics are identified: number of classes, number of instances, number of attribute, class entropy, class skew, and class conditional entropy by columns.

These characteristics are linked to different feature selection approaches under different search techniques, along with applying six classifiers (J48, naiveBayes, oneR, part, zeroR, and neural network) and fed to J48 meta learner in order to create a recommendation on what feature selection approaches should be used on a specific data set. TABLE V summarizes the feature selection approaches, evaluator strategies, and the search methods that are used in our experiment. None indicates using the learning algorithm without making any feature selection.

TABLE V. COMBINATION OF FEATURE SELECTION APPROCHES, SEARCH STRATEGIES AND EVALUATORS

Feature selection approach	Attribute evaluator	Search strategy
None	None	None
Wrapper	ClassifierSubSetEval	GreedySearch
Wrapper	ClassifierSuSetEval	BestFirt
Filter	InfoGain	Ranker
Filter	cfsSubSetEval	Ranker
Filter	GainRatioRval	Ranker

A sample of the data characteristics with different feature selection approaches along with the different classifier accuracy and cost and for contact-lenses data set is shown in Table VI.

To obtain the data on performance over different feature selection methods, desired features are extracted for 26 data sets , each data set with its characteristics is linked to 5 classifiers to evaluate their performance and cost after applying 6 different feature selection combination that listed in TABLE V, so each data set has 30 rows (5 classifier * 6 feature selection combination) and linked to 6 data set characters listed in TABLE VI , then the result is fed to J48 learner using 10 folds cross validation

The following is the decision trees result from applying j48 as a meta learner on the previous data using 10 folds cross validation, Fig. A-1,A-2 (APPENDIX A) show the decision tree to predict the classifier performance (accuracy and cost) .

TABLE VI. SAMPLE OF DATASET CHARATARISTICS USED TO BUILD A META LEVEL LEARNER ALONG WITH CLASSIFIER PERFORMANCE

number of classes	number of attribute	Class entropy	Number of instances	Class conditional entropy	Class skew
3	5	1.3	24	3.7	0
Feature Selection	**Type**	**Search Method**	**Classifier**	**Accuracy**	
None	None	None	J48	80-85	
			NaiveBayes	70-75	
			OneR	70-75	
			Part	60-65	
			Neural Network	70-75	
Feature Selection	**Type**	**Search Method**	**classifier**	**Cost**	
None	None	None	J48	30-40	
None	None	None	NaiveBayes	50-60	
None	None	None	oneR	50-60	
None	None	None	Part	30-40	
None	None	None	Neural Network	50-60	

The knowledge in Figure is suitable as meta knowledge that guides the data mining process through its journey, consider

a specific data set is given, all data characters for this data set are allocated and the performance of all target learners are predicted using the previous decision tree with best feature selection strategy and approach (Figs. A-1, A-2) shown in APPENDIX A , for example if a new data set is emerged, the data characters are automatically calculated so if the class entropy is greater than 0.28, the class skew is <-0.08, number of attribute >28 , if the user uses a naiveBayes classifier, so the predicated performance is 80-85 if oneAttributeEval feature selection is used. TABLE VII shows part of the meta knowledge output for credit-g data set [24].

TABLE VII. THE META KNOWELDGE OUTPUT

Data Set	Classifier	Accuracy	Feature Selection	Cost	Cost Feature Selection
credit-g	J48	70-75	None	80-85	None
	NaiveBayes	80-85	None	80-85	None
	oneR	70-75	None	>100	None
	Part	70-75	Wrapper	80-85	None
	Part	70-75	Filter	90-95	None
	Part	70-75	None	80-85	None
	ZeroR	66-70	None	>100	None
	Neural Network	70-75	None	70-75	None
	Neural Network	70-75	None	60-65	GreedySearch
	Neural Network	70-75	None	60-65	BestFirst
	Neural Network	70-75	None	70-75	Ranker

VI. CONCLUSION

Feature selection is a significant step in data mining and many feature selection methods and search strategies have been developed. No single method and strategy are dominant and hence a data miner has to spend time experimenting in order to determine the most appropriate feature selection methods to use for a particular data set.

Hence this paper develops a meta learning framework for learning from the experience of applying difference feature selection methods. The framework has been developed in the Weka toolkit and the results are presented. These results are promising and show that the meta-knowledge produced appears useful.

The next stage of the research involves experiments to see how well it evolves with more data sets and to include concepts of active learning to make it scalable.

REFERENCES

[1] H,Almuallim., &Dietterich, T. G. “Efficient algorithms for identifying relevant features”. In Proc. of the 9th Canadian Conference on Artificial Intelligence, pp. 38-45.(1991).

[2] Kira, Kenji, and Larry A. Rendell. "The feature selection problem: Traditional methods and a new algorithm." In AAAI, pp. 129-134. 1992.

[3] John, George H., Ron Kohavi, and Karl Pfleger. "Irrelevant Features and the Subset Selection Problem." In ICML, vol. 94, pp. 121-129. 1994..

[4] Kohavi, Ron, and George H. John. "Wrappers for feature subset selection." Artificial intelligence 97.1 : p.p.273-324.(1997)

[5] Das, Sanmay. "Filters, wrappers and a boosting-based hybrid for feature selection." In ICML, vol. 1, pp. 74-81. 2001

[6] Koller, Daphne, and MehranSahami. "Toward optimal feature selection." (1996).

[7] Dash, Manoranjan, Kiseok Choi, Peter Scheuermann, and Huan Liu. "Feature selection for clustering-a filter solution." In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, pp. 115-122. IEEE, (2002).

[8] Jensen, Richard, and Qiang Shen. "New approaches to fuzzy-rough feature selection." Fuzzy Systems, IEEE Transactions on 17.4: 824-838.(2009) .

[9] Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." Evolutionary Computation, IEEE Transactions on 1.1 :67-82. (1997).

[10] Liu, Huan, and Hiroshi Motoda. Feature selection for knowledge discovery and data mining. Springer, (1998).

[11] Miller, Alan. Subset selection in regression. CRC Press, (2002).

[12] Molina, Luis Carlos, LluísBelanche, and ÀngelaNebot. "Feature selection algorithms: A survey and experimental evaluation." In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, pp. 306-313. IEEE,(2002).

[13] Wang, Juan, Lin-lin Ci, and Kan-Ze Yao. "A survey of feature selection." Computer Engineering and Science 27, no. 12 :68-71.(2005).

[14] Aha, David W., and Richard L. Bankert. "A comparative evaluation of sequential feature selection algorithms." Learning from Data. Springer New York,. 199-206.(1996).

[15] Skalak, David B. "Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms." In ICML, pp. 293-301. (1994).

[16] Yang, Jihoon, and VasantHonavar. "Feature subset selection using a genetic algorithm." In Feature extraction, construction and selection, pp. 117-136. Springer US, (1998).

[17] Vasantha, M., V. SubbiahBharathi, and S. Dhamodharan. "Medical image feature, extraction, selection and classification." International Journal of Engineering Science 2 .(2010).

[18] Liu, Y., and M. Schumann. "Data mining feature selection for credit scoring models." Journal of the Operational Research Society 56.9 p.p 1099-1108.(2005).

[19] Shilbayeh, Samar,and Vadera,Sunil.”A meta learner framework based on landmarking”. European Conference on Intelligent Management Systems in Operations (IMSIO2013) Operational Research Society, p.p.97-105. (2013).

[20] Ron Kohavi, George H. John .Wrappers for feature subset selection. Artificial Intelligence. 97(1-2):273-324. (1997).

[21] M. A. Hall. Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand. (1998)

[22] R. Jensen, Q. Shen. New Approaches to Fuzzy-rough Feature Selection. IEEE Transactions on Fuzzy Systems. (2009).

[23] X. Wang, J. Yang, X. Teng, W. Xia,, R. Jensen. Feature Selection based on Rough Sets and Particle Swarm Optimization. Computer Methods and Programs in Biomedicine. 83(2):459-471. (2007)

[24] Asuncion, Arthur, and David Newman. "UCI machine learning repository." (2007).

APPENDIX A

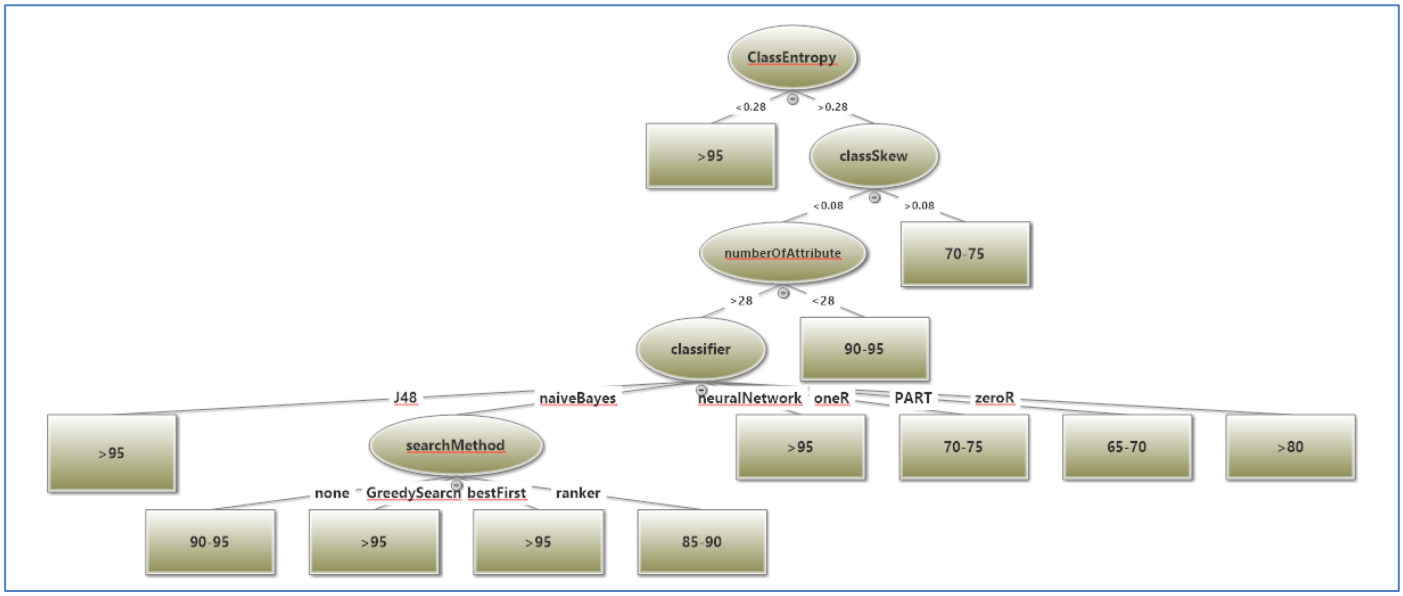

Fig A-1 Classifier performance prediction meta knowledge

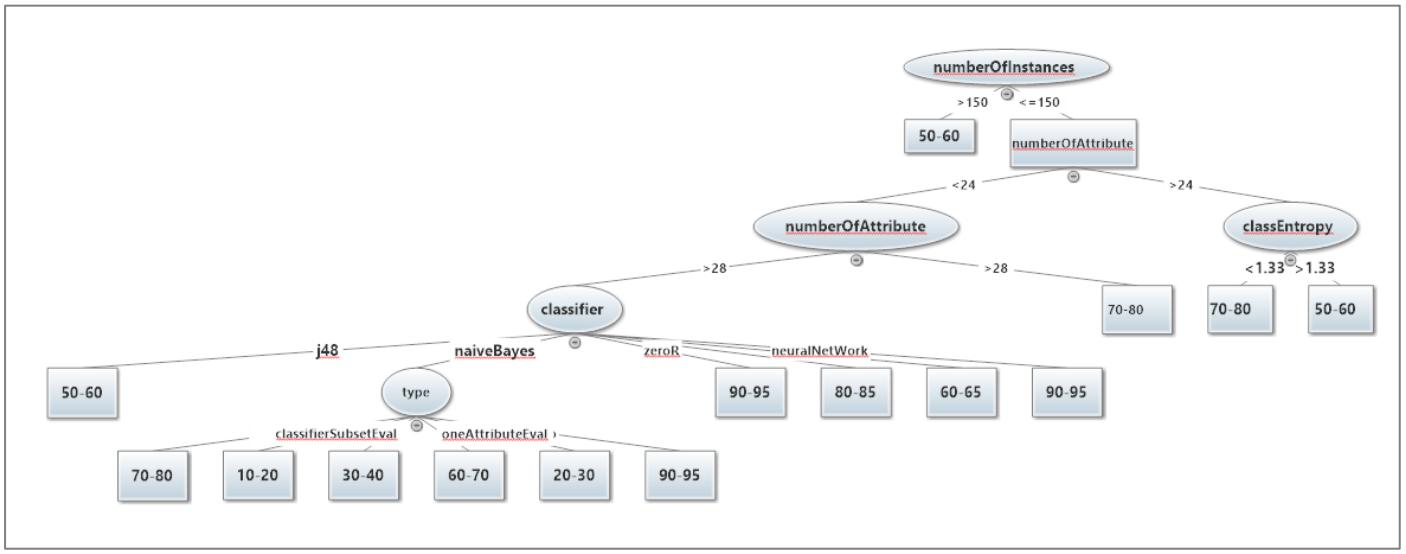

Fig A-2 Classifier cost prediction meta knowledge