International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Aspect Range Motivated Classifier Ensemble Reduction

Ganeshkumar .S[1], P. Selvaraj[2]

[1]M. Tech-Information Technology, SRM University, Chennai, Tamilnadu, India

[2]Assistant Professor, Department of Information Technology, SRM University, Chennai, Tamilnadu, India

Abstract: ***This paper proposed about classifier which ensemble in data mining and it constitute one of the directions in main research of machine learning. In general multiple classifiers allow better predictable performance. To construct and aggregate ensembles there are several approaches that exists in the literature. To produce better results and to increase group diversity redundant members should be removed which contain in the ensemble systems. Smaller ensembles helps to relax storage requirements and the memory which improves efficiency by reducing run time overhead of the system. In this paper the ideas are extended for development of feature selection problems by transformation of training samples from ensemble predictions which supports reduction of classifier ensembles. To maximize the evaluation of feature subset from the selection of reduced subset of artificial feature this project use global heuristic harmony search. The large sized and high dimensional benchmark datasets are used to evaluate the resulting technique systematically it shows superior performance of classifications against randomly formed subsets and unreduced, original ensembles.***

Keywords: Harmony search, classifier ensemble, HSFS (Harmony Search Feature Selection), CER (Classifier Ensemble Reduction);

1. Introduction

In a range of sequential data mining applications, a classifier ensemble is to boost the performance of single classifier systems. Totally different classifiers sometimes build different predictions on sure samples, caused by their various internal models. Combining such classifiers has become the natural manner of attempting to extend the classification accuracy, by exploiting their unrelated errors. Also, every ensemble member can potentially be trained employing a subset of coaching samples, which can scale back the complex computational process quality issue that arises only one classification algorithmic rule is applied to terribly massive knowledge data sets. In additionally, an ensemble will operate in a very distributed area, wherever datasets are physically separated and are backgrounds before combining their selections along to supply the ultimate prediction. Classifier Ensemble Reduction (CER) is to cut back the quantity of redundancy in a very pre-constructed classifier ensemble, to make a way to reduced subset of classifiers that may be still deliver constant classification results. Having a reduced range of classifiers will eliminate some of run-time overheads, creating the ensemble process faster than having fewer models additionally suggests that relaxed memory and storage needs. Removing redundant ensemble members can also cause improved diversity among the cluster, and more increase the prediction accuracy of the ensemble. Existing approaches within the literature include techniques that use clustering to find groups of models that share similar predictions, and after prune every cluster singly. Others use reinforcement learning and multi label learning to attain redundancy removal. Variety of comparable approaches specialize in choosing an optimal best subset of classifiers, to maximize an explicit predefined diversity live. Inspired by the analogies in between CER and FS, this approach tries to find a set of classifiers by eliminating redundant cluster members, whereas maintaining the quantity of diversity among the initial ensemble. Harmony Search (HS) acts as a metaheuristic algorithmic rule that tries to search out an answer vector, that optimizes a given price operate. HS-based FS algorithmic rule that is the elemental platform upon that the CER system is developed. The HS-based FS technique (HSFS), and explains however FS downside may be change into an improvement problem, more resolved by HS.

2. Existing System

Different classifiers sometimes build totally different predictions on sure samples, caused by their different internal models. Combining such classifiers has become the natural manner of attempting to extend the classification accuracy, by exploiting their unrelated errors. Also, every ensemble member will be potentially be trained employing a set of coaching samples, that could reduce the procedure complexness issue that arises only one classification algorithmic rule is applied to terribly massive datasets. In additionally, an ensemble will operate in a very distributed area, wherever datasets are physically separated and are price ineffective or technically tough to be combined into one database, to train a single classifier. Rather than adopting an easy majority voting-based aggregation, ways have conjointly been developed that use meta-level learners to mix the outputs of the base classifiers. Existing techniques that use clustering to find groups of models that share similar predictions, and afterward prune every cluster severally. Variety of comparable approaches concentrate on choosing a potentially optimum subset of classifiers, to maximize a definite predefined diversity measure. Optimality is subjective depending on the problem at hand, and a subset that's chosen as optimum using one explicit authority evaluator might not be equivalent to that of a subset chosen by another. An unsupervised FS methodology that operates on unlabeled data. In distinction, wrapper-based and hybrid algorithmic rules are typically utilized in conjunction with a learning or data processing algorithm that is utilized in situ of an analysis metric as utilized in the filter-based approach. Hill-climbing-based approaches are exploited wherever options are additional or removed one at a time till there's no more improvement to this candidate answer. Dynamic parameter standardization and iterative solution refinement techniques have conjointly been planned to more improve the search outcome.

Disadvantages
1. It won't access too much of data from the dataset for a particular details.
2. It shows too much of unreduced irrelevant data for the required.
3. System overhead by using single classifier instead of multi classifier.
4. It uses more memory space to access and perform some tasks.

3. Proposed System

A new framework for CER is proposed that builds upon the concepts from existing FS techniques. Intensify by the analogies in between CER and FS, this approach makes an attempt to find a subset of classifiers by eliminating redundant cluster members, whereas maintaining (or increasing) the number of diversity among the initial ensemble. As a result, the CER problem is being tackled from a different angle. Every ensemble member is currently transformed into an artificial feature during a new created dataset, and also the feature values are generated by aggregation the classifiers predictions. FS algorithms will then be used to remove redundant options within the present context, to pick out a smallest classifier subset whereas maintaining original ensemble diversity, and conserving ensemble prediction accuracy. The target of classifier ensemble reduction (CER) (or classifier ensemble pruning) is to scale back the number of redundancy during a reconstructed classifier ensemble, to create a lot of reduced subset of classifiers which will still deliver the same classification results. It's associate intermediate step between ensemble construction and decision aggregation. Efficiency is one amongst the obvious gains from CER. Having a reduced range of classifiers will eliminate a portion of run-time overheads, creating the ensemble process quicker; removing redundant ensemble members can also cause improved diversity among the cluster, and any increase the prediction accuracy of the ensemble. The most aim of feature selection (FS) is to find a smallest feature subset from a problem domain whereas holding a suitably high accuracy in representing the initial information. HS acts as a Meta heuristic rule that makes an attempt to search out a results vector that optimizes a given (possibly multivariate) value operation. In such a research method, every decision variable (musician) generates a value (note) for locating a global optimum (best harmony).

Advantages
- The new scheme guarantees high data privacy.
- Provide heavy security for storage.
- Error detection and Error correction to reduce irrelevant data to increase memory efficiency.

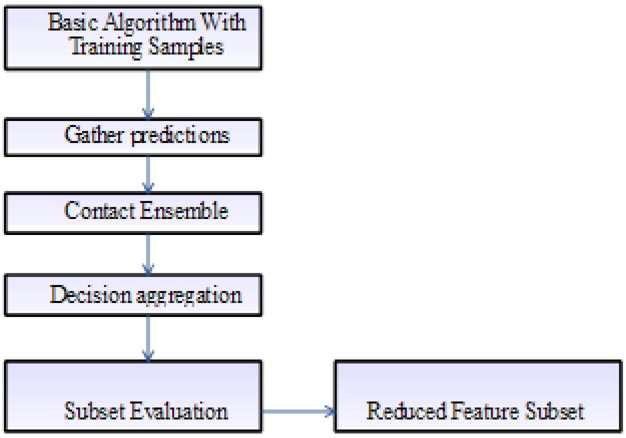
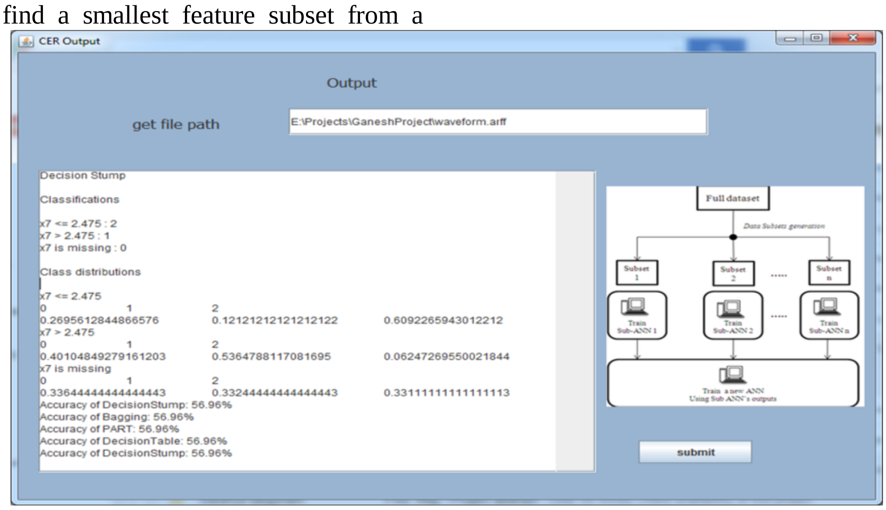


Figure 1: Architecture Diagram

In this the classifier ensemble is generated and trained using a set of given training data. For new samples, each ensemble member individually predicts a class label, which are aggregated to provide the ensemble decision. It is inevitable that such ensembles contain redundant classifiers that share very similar if not identical models. Classifiers is to maintain and improve the ensemble diversity. The fundamental concept and goals of CER is therefore the same as FS. Having introduced the HSFS technique, the following section aims to explain how a CER problem can be converted into an FS scenario, and details the framework proposed to efficiently perform the reduction.

4. CER Framework



In this frame work I select this particular dataset like weather reports from UCI web we can get different set of data can access and compare in this CER to increase resulting accuracy. It can do the decision matrix method by default to partition the dataset to reduce the irrelevant member in classifier ensembles. Globally increases memory efficiency and avoid to predict the irrelevant data. In this it uses the following modules to perform the operations to produce better predictive results.

5. Modules Details

1. Base Classifier Pool Generation.
2. Classifier Decision Transformation.
3. Feature Selection on the Transformed Data Set.
4. Ensemble Decision Aggregation.

1) Base Classifier Pool Generation

Base Classifier Pool (BCP) is that the initiative process in producing sensible CER. Any most popular classifier is used to build in the base classifier, like bagging or bootstrap aggregation could be a machine learning ensemble designed to enhance the stability and accuracy of machine learning algorithm used to applied mathematics classification & regression. During this base algorithm and training samples is combined and generate bagging it choose completely different subset from dataset to create various classifier.

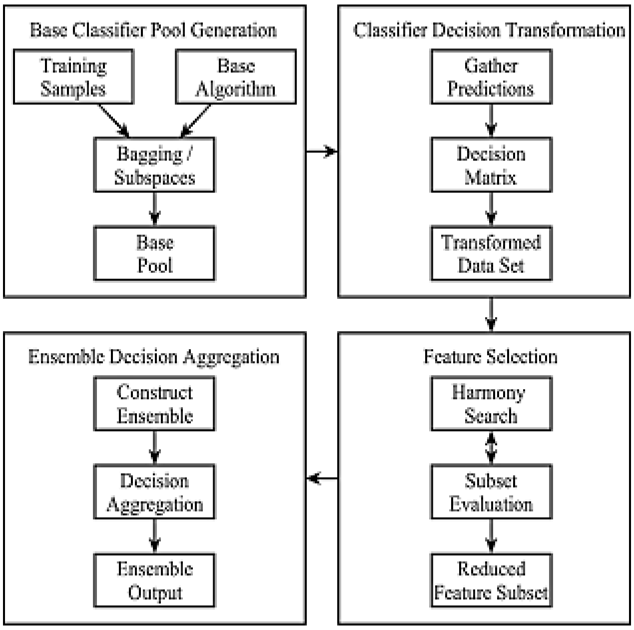


Figure 2: CER diagram

2) Classifier Decision Transformation

Once the base classifiers area unit designed, their choices on the training instances are also gathered. For base classifiers Ci, i =1, 2, . . . ,NC, and training instances Ij, j = 1, 2, . . . ,NI ,where NC is that the total number of base classifiers, and Ni is the total number of training instances, a decision matrix as shown in Table I is created. The worth Dij represents the ith classifier's decision on the jth instance.

Table 1

$$
\begin{array}{c|ccccc}
 & C_1 & C_2 & \cdots & C_i & \cdots & C_{N_C} \\
\hline
I_1 & D_{11} & D_{21} & \cdots & D_{i1} & \cdots & D_{N_C 1} \\
I_2 & D_{12} & D_{22} & \cdots & D_{i2} & \cdots & D_{N_C 2} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
I_j & D_{1j} & D_{2j} & \cdots & D_{ij} & \cdots & D_{N_C j} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
I_{N_I} & D_{1N_I} & D_{2N_I} & \cdots & D_{iN_I} & \cdots & D_{N_C N_I}
\end{array}
$$

A decision matrix is list of values in rows associated columns that permits an analyst to consistently identity, analyse and rate the performance of relationship between sets of values and data. A new dataset is therefore created, every column represents associate artificially generated feature, and every row corresponds to a coaching instance, the cell then stores the transformed feature value.

3) Feature Selection on the Transformed Data Set

HSFS is then performed on the artificial dataset, evaluating the rising feature subset exploitation the predefined subset evaluator. HSFS optimizes the standard of discovered subsets, while trying to reduce subset sizes. Once HS terminates, its best harmony is translated into a feature set and return as the FS result. The options then indicate their corresponding classifiers that should be enclosed within the learnt classifier ensemble.

4) Ensemble Decision Aggregation

Once the classifier ensemble is constructed, new objects are classified by the ensemble members, and their results are aggregated to form the final ensemble decision output. The average of probability method is used in this paper. Given ensemble members Ei, i = 1, 2, . . . ,NE, and decision classes Dj, j = 1, 2, . . . , ND, where NE is the ensemble size and ND is the number of decision classes, classifier decisions can be viewed as a matrix of probability distributions {Pij}. Here, Pij indicates prediction from classifier Ci for decision class Dj.

$$\{\sum_{i=1}^{N_E} P_{i1}/N_E, \sum_{i=1}^{N_E} P_{i2}/N_E, \ldots, \sum_{i=1}^{N_E} P_{iN_D}/N_E\}.$$

The final aggregated decision is the winning classifier that has the highest averaged prediction across all classifiers, As such, the same old different aggregation methodology majority vote is not any longer favourable since the bulk has now been considerably reduced.

Table II

HMS	#Musicians	HMCR	K
10-20	*# features*	*0.5-1*	*1000*
Data set	*Features*	*Instances*	*Decisions*
Arrhythmia	*280*	*452*	*16*
Magic	*10*	*19020*	*2*
Waveform	*41*	*5000*	*3*

A collection of real-valued UCI benchmark datasets are used in the experiments, a number of which are large in size and high in dimension; hence, present significant challenges to the construction and reduction of ensembles. The parameters used in the experiments and the information of the datasets are summarized in Table II. Ratified ten-fold cross-validation is employed for data validation. The construction of the base classifier ensemble, and the ensemble reduction process are both performed using the same training fold, so that the reduced subset of classifiers can be compared using the same unseen testing data. The stratification of the info its division into completely different folds ensures that every category label has equal illustration all folds, thereby serving to alleviate bias/variance issues. The experimental outcomes presented are averaged values of 10 completely different 10-FCV runs (i.e., a hundred outcomes), to minimize the impact of random factors within the heuristic algorithms.

6. Conclusion

This project ensures to search data using CER to reduce the irrelevant members from the dataset. I develop this project to increase the resulting factor to more accuracy. In the feature it predict to show the weather forecasting details can be already collected and stored as dataset and we can run this dataset in CER and it will predict the feature result. It also useful for predicting the heart disease result in perfect by already stored heart beat rhythm and as per the heart beat rhythm it calculate the accurate results. It also produce the better predictive results.

References

[1] G. Tsoumakas, I. Partalas, and I. Vlahavas, "A taxonomy and short review of ensemble selection," in *Proc. Workshop Supervised Unsupervised Ensemble Meth. Their Applicat.*, 2008,

[2] K. S. Lee and Z. W. Geem, "A new meta-heuristic algorithm for continuous engineering optimization: Harmony search theory and practice," *Comput. Meth. Appl. Mech. Eng., vol. 194, nos. 36–38, pp. 3902–3933,* Sep. 2005.

[3] L. Kuncheva, "Switching between selection and fusion in combining classifiers: An experiment," *IEEE Trans. Syst., Man, Cybern. B, Cybern,* vol. 32, no. 2, pp. 146–156, Apr. 2002.

[4] M. A. Tahir, J. Kittler, and A. Bouridane, "Multilabel classification using heterogeneous ensemble of multi-label classifiers," *Pattern Recognit.2012*

[5] N. Mac Parthal´ain and R. Jensen, "Measures for unsupervised fuzzyrough feature selection," Int. J. Hybrid Intell. Syst., vol. 7, no. 4, pp. 249–259, Dec. 2010.

[6] R. W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recognit. Lett.2003*

[7] S.-M. Zhou and J. Q. Gan, "Constructing accurate and parsimonious fuzzy models with distinguishable fuzzy sets based on an entropy measure," *Fuzzy Sets Syst.*,2008

[8] X. Li and L. Parker, "Design and performance improvements for fault detection in tightly-coupled multi-robot team tasks," in *Proc. IEEE Int. Conf. Robot. Autom., Apr. 2008, pp. 198–203.*

[9] Z. Zhu, Y.-S. Ong, and M. Dash, "Wrapper-filter feature selection algorithm using a memetic framework," *IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 37, no. 1, pp. 70–76, Feb. 2007.*