International Journal of Engineering & Technology, 7 (2.33) (2018) 235-238

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A novel approach to ensemble learning in distributed data mining

Prakash Chandra Jena [1] *, Subhendu Kumar Pani [2], Debahuti Mishra [1]

[1] Siksha O Anusandhan Deemed to be University, Odisha, India
[2] Orissa Engineering College, Odisha, India
**Corresponding author E-mail: prakashjena81@gmail.com*

Abstract

Several data mining techniques have been proposed to take out hidden information from databases. Data mining and knowledge extraction becomes challenging when data is massive, distributed and heterogeneous. Classification is an extensively applied task in data mining for prediction. Huge numbers of machine learning techniques have been developed for the purpose. Ensemble learning merges multiple base classifiers to improve the performance of individual classification algorithms. In particular, ensemble learning plays a significant role in distributed data mining. So, study of ensemble learning is crucial in order to apply it in real-world data mining problems. We propose a technique to construct ensemble of classifiers and study its performance using popular learning techniques on a range of publicly available datasets from biomedical domain.

Keywords: *Ensemble Learning; Meta-Learning; Classifier Ensemble; Ensemble Method; Classification Performance; Meta-Classifier.*

1. Introduction

Due to the advancement of computing and communication technology over wired and wireless network have outcome in many pervasive distributed computing field. There has been an explosive growth of data available in many of these environments. It leads to a large-scale data analysis problem and offers an opportunity to develop automated data mining techniques for discovering patterns in the massive data and extracting essential knowledge from it. The difficulty of data mining is further aggravated due to the fact that in many cases, the data is distributed over many computing nodes and remain heterogeneous. Data distribution can be accounted to several factors such as ownership and privacy. For example, several datasets concerning crucial business information (e.g. credit card fraud, money laundering) might be owned by different organizations located geographically in several locations and they have genuine reasons to keep the data private. However, they may be interested in sharing these data for useful information and better interest of the business. Therefore, the issues of modern data mining techniques include not just the size of the data to be mined but also its distribution and heterogeneity.

A widely adopted approach to the distributed data mining problem is to apply various machine learning algorithms using parallel and incremental learning techniques. In particular, meta-learning approach has been widely used in many studies [1-2]. Meta-learning involves ensemble learning wherein several learning techniques act as core learners and develop local models on distributed data sources. A higher-level learning algorithm that produces a final model of the distributed data then combines these local models. A general framework of ensemble learning in distributed data mining is shown in Figure-1.

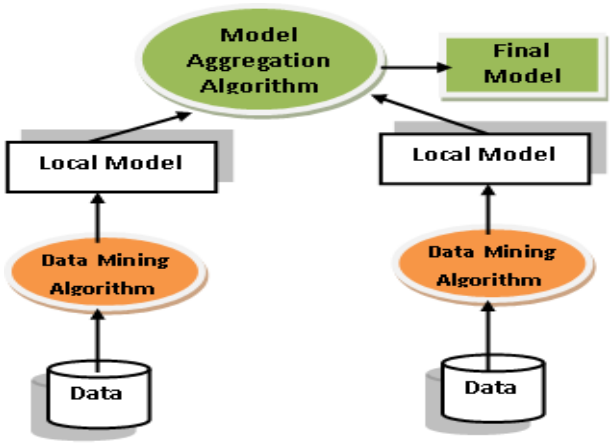

Fig. 1: Ensemble Learning Framework

Ensemble learning techniques tries towards progressing the high expectation of a learning system throughout incremental learning. Studies claim that ensemble learning performs better accuracy than that of basic core learners while eliminating biasness of a particular algorithm and keeping unambiguousness[3]. Numerous ensembles based learning methods such as boosting, bagging, voting and stacking have been implemented in the literature of machine learning applications [4-5].

Considering the result of ensemble learning is significant in order to use it to real-world problems6. As a result it has been a motivating topic of research for data mining community over the years[7]. In this paper, we investigate into ensemble learning, develop a method of constructing ensemble of classifiers, and report how it affects the classification performance in comparison to individual learning methods.

The rest of the paper is prepared as follows. Reviews of the related works on ensemble learning are presented in section-2. Section 3

describes our methodical approach adopted for the study. Section 4 presents the findings of this investigation and then section 5 concludes the paper.

2. Related work

Ensemble learning has widely studied across several application domains and is evident in the literature due to the fact that ensembles can often perform better than any single classifier [8-9].

Liu et al. (2012) 10 created huge hierarchical based ensembles techniques of many fold classifiers and applied them for the analysis of Alzheimer's disease. A MATLAB program separates the entire brain figure into a number of neighboring 3D areas and instructs two classifier of low-level for each area. Then it constructs a new set of advanced high-level classifiers corresponding to different areas of the brain and chooses a subset of the high-level classifiers using a forward greedy search strategy. Lastly, it merges the result of the chosen advanced high-level classifiers using a weighted voting.

Islam and Abawajy (2013)11 proposed a multi-level model for classification in phishing email filtering. The approach extracts the characteristics of fraudulent emails derived from the content of the message and header of the message and selects characteristics according to a priority position. An n-tier classification process is used to detect and filter phishing emails.

Xiao et al. (2012)12-studied combination of cost sensitive learning with ensemble learning to deal with class imbalance problem. It uses cost sensitive selection criteria for Dynamic Ensemble Selection (DES) and Dynamic Classifier Selection (DCS) to enhance the classification capability for imbalanced data.

Kelarev et al. (2012)13 investigated ensemble learning using meta-classifiers in multi-level for the classification of cardiac neuropathy progression. It uses a large number of base classifiers and several meta-classifiers to study classification performance. The outputs show that meta-classifiers and multi-level ensemble meta-classifiers can be applied to improve the classifications accuracy further.

Fumera et al. (2005)14 proposed a conceptual and experimental study of linear combination for fusion of classifier. Their conceptual study shows how the results of linear combiners depend on the result of particular classifiers, and on the correlation among their results. Especially, they measured the improved results obtained from appling a weighted average over the plain average-combining rule.

Al-Razgan and Domeniconi (2009)15 used clustering with the help of ensemble, and address the difficulty of joining multiple biased clusters that comes from diverse subspaces of the input space. They leverage the variety of the input clustering in order to produce a consensus partition that is higher to the participating ones. Their answers were as good as or improved than the best particular clustering, providing the entered input clustering was varied.

3. Our approach

We implement an experimental machine learning approach to perform the study which is explained in the next subsections.

3.1. Datasets used

The applied datasets in the experiment are taken from widely existing UCI Machine Learning Repository. The sample is rigorously chosen to replicate diversity with a varied group of
statistical characteristics biomedical relevance domain. Table-1 shows the properties of the datasets.

Table 1: Characteristics Dataset

Dataset	#No. of Instances	#No. of Classes	#No. of Attributes	#Nominal / Continuous	#Missing Values (%)
Breast-Cancer	286	2	9	9/0	0
Diabetes	768	2	8	0/8	0
Heart-statlog	270	2	13	0/13	0

3.2. Classifiers used

A group of four widely used classification algorithms is selected for base-level learning. They are Sequential Minimal Optimization (SMO), Naïve Bayes (NB), Bike and J48.

The Bayesian Network based Naïve Bayes (NB) is an approach that applies statistical techniques to classify attributes derived from probabilities. It is very robust to isolated noise points, inappropriate properties and manages missed values. Sequential Minimal Optimization (SMO) is an advanced technique based on Support Vector Machine and is very well organized for optimization case. It performs well with sparse data. K-Nearest Neighbor technique i.e. IBk that outputs databased on instance-based methodical learning. It presents better performance when number of instances is very high. J48 is an execution of higher C4.5 techniques. It builds decision tree from a group of training data based on information entropy. Every node of the tree symbolizes the most efficient divide of the samples determined by the maximum normalized information gain. It is able to handle both numeric and nominal attributes with missing value attributes. It executes well with nominal attributes [16].

3.3. Ensemble selection

A set of three popular meta-classifiers is selected for ensemble learning and comparative evaluation. They are Decorate, AdaBoost and Bagging. A detailed description of each of these techniques is provided here.

Decorate stands for varied Ensemble Creation by Oppositional Relabeling of Artificial Training Examples. Decorate can use any base classifier and builds ensemble of classifiers. Decorate builds unique artificial training examples to create diverse hypotheses for constructing diverse ensembles of classifiers. Decorate consistently generates ensembles more correct than the base classifier [17].

AdaBoost, boosting employs numerous classifiers in series. Every classifier is employed on the instances that contain cast out harder for the previous classifier. To solve this, all instances are allocated weights, and if an instance changes hard to classify, then its weight rises [18].

Bagging produces a group of fresh sets by re-sampling the existing training set at random and with substitution. These collections are known as bootstrap samples. Fresh classifiers are then trained, one for every new training sets. They are compounded via a popular vote [5].

3.4. Weka spark

Weka has been used to conduct the experiment. It is a comprehensive extensively used java based data mining tool with well-known machine learning algorithms and intuitive interface supporting all phases of the data mining stages. However, it performs only sequential single node execution. To enable distributed data mining process and scalable processing, Weka Spark has been combined with sequential Weka. Weka Spark is a distributed framework implemented on top of Spark that provides quick in-memory processing capabilities and maintain for iterative computations. Weka Spark leverages Weka's ability and Spark's distributed capability enabling sequential Weka for distributed simulation [19].

3.5 Experiment design

Our experiment evaluates three meta-classifiers for multi-level ensemble learning that considers four base classifiers. However, we restrict ensemble construction to two-level in order to reduce the computational complexity. The single ensemble learning is constructed using a meta-classifier of level 1 and a base classifier (for example, Decorate + J48) while the layered ensemble learning is constructed using a meta-classifier of level 2, a meta-classifier of level-1 and a base classifier (for example, Decorate + Bagging + J48).

We perform the experiment using cross validation of 10 fold as the test mode to trace classification accuracy. The 10-fold cross validation avoids influenced results and provides strength to the categorization. Further, the attributes of classification methods are selected to their default values. The subsequent steps are used to construct classifier ensemble and study the performance.

Step-1: Run all candidate classification algorithms one by one on each of the datasets to trace each one's classification accuracy using both Weka and Weka Spark.

Step-2: Select the classifiers which provide consistently better accuracy across the datasets for Weka Spark environment over Weka. These classifiers are assumed to be base classifiers for ensemble learning.

Step-3: Formulate dataset scenarios by using combinations of chosen base classifiers with all the datasets.

Step-4: Run all the ensemble algorithms on the formulated dataset scenarios to record classification accuracy of the single ensemble approach in Weka Spark environment.

Step-5: Formulate 2-layered ensemble using combinations of ensemble algorithms.

Step-6: Run each of the 2-layered ensembles on the dataset scenarios to record the classification performance in Weka Spark environment.

The configuration of Weka Knowledge Flow Environment using Weka Spark to run classifiers is shown in Figure-2.

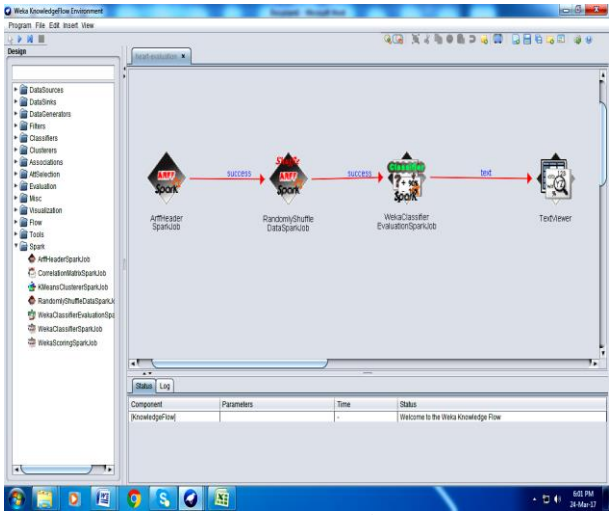

Fig. 2: Weka Spark Configuration.

4. Result analysis

The accuracy in classification of the candidate classifiers on the chosen datasets after applying step-1 of the experiment are shown in Figure-3. Two classifiers i.e. NB and SMO are knocked out in the step-2 of our experiment as they fail to perform consistently across the datasets. Only IBk and J48 are considered for subsequent stages of the experimental study.

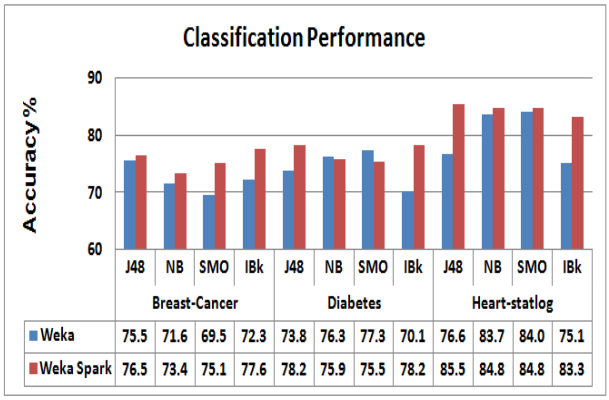

Fig. 3: Performance of Candidate Classifiers.

The dataset scenarios formulated using these two classifiers for ensemble learning are shown in Table-2.

Table 2: Dataset Scenarios

Dataset	Classifier	Scenario
Breast-Cancer	J48	BCJ48
	IBk	BCIBk
Diabetes	J48	DJ48
	IBk	DIBk
Heart-statlog	J48	HSJ48
	IBk	HSIBk

The accuracy percentage of single ensemble learning against selected base classifiers in different dataset scenarios is shown in Table-3 and performance of single ensemble learning in terms of percentage of improvement in accuracy is depicted in Figure-4. It reveals that classification accuracy of Decorate ensemble is consistently significant across the scenarios. In addition, AdaBoost provides improved accuracy than that of the base classifier except one scenario (i.e. BCJ48) wherein there is no change in accuracy. However, bagging algorithm has inconsistent performance.

Table 3: Accuracy of Single Ensemble vs. Base Classifier

Dataset Scenario	Decorate	AdaBoost	Bagging	Base Classifier
BCJ48	79.7203	76.5734	75.5245	76.5734
BCIBk	83.2168	80.0699	80.4196	77.6224
DJ48	79.9479	79.0365	81.25	78.2252
DIBk	84.5052	79.5573	78.6458	78.2552
HSJ48	88.5185	87.7778	84.8148	85.5556
HSIBk	90	83.7037	85.5556	83.3333

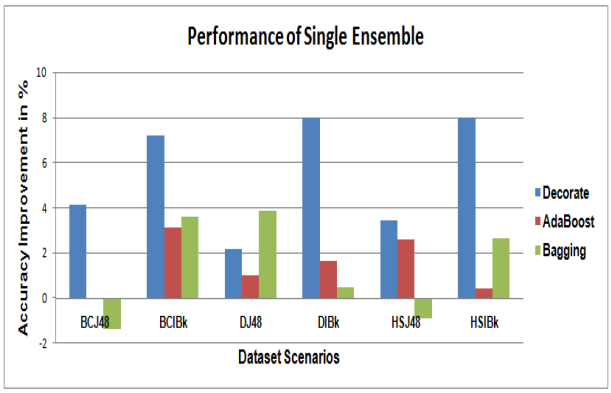

Fig. 4: Accuracy Improvement of Single Ensemble.

The accuracy percentage of layered ensemble learning against base classifiers on different dataset scenarios is shown in Table-4 and performance of single ensemble learning in terms of percentage of improvement in accuracy is depicted in Figure-5. It shows that ensemble learning with Decorate + AdaBoost combination provides significantly better performance and consistent across the scenarios. It also shows that ensemble learning with Decorate + Bagging combination is somewhat better and consistent across the scenarios. However, other combinations such as AdaBoost + Dec-

orate, Bagging + Decorate, AdaBoost + Bagging and Bagging + AdaBoost show negative performance. It can be established from this experimental data that Decorate when placed as level-2 meta-classifier performs well consistently across the datasets considered in the study and improves the classification accuracy between two to eleven percent (i.e. 2% - 11%) as compared to the accuracy achieved by the base level classifier.

Table 4: Accuracy of Layered Ensemble vs. Base Classifier

Dataset Scenario	Decorate and AdaBoost	Decorate and Bagging	Bagging + Decorate	Bagging andAda-Boost	AdaBoost andDecorate	AdaBoost andBagging	Base Classifier
BCJ48	84.965	82.8671	81.1181	80.4196	78.6713	77.6224	76.5734
BCIBk	84.6154	80.7692	82.1678	74.8252	77.2727	78.3217	77.6224
DJ48	83.4635	84.7656	82.0313	81.5104	81.1198	80.9896	78.0365
DIBk	84.5052	80.7292	81.3802	78.6458	76.1719	78.125	78.2552
HSJ48	90.7407	87.4079	85.5556	87.4074	84.4444	86.2963	85.7778
HSIBk	90	88.5185	90.3704	85.5556	83.3333	85.1852	83.3333

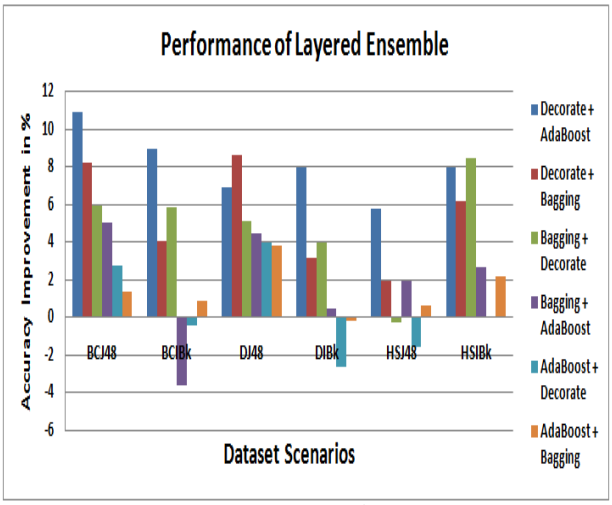

Fig. 5: Accuracy Improvement of Layered Ensemble.

5. Conclusion

In this study, we selected two base classifier out of four considered initially. The selected base classifiers combined with popular meta-classifiers in multi-level ensemble learning on datasets taken from biomedical domain. Simulations were carried out using Weka Spark distributed framework. Based on the data generated it is concluded that Decorate algorithm performs very well and provides significant classification accuracy as compared to the classification accuracy achieved by the individual base classifiers. It improves its classification performance further across the datasets when placed as a level-2 classifier in the multi-level ensemble construction. However, its classification performance becomes inconsistent when it is placed as lavel-1 meta-classifier in the ensemble. Nevertheless Decorate performs well and is the most excellent option for all selected datasets. Additional it can be thought that meta-learning gets better classification accuracy over base-learning and the experiential data used in this study provides a better result to it.

References

[1] Vilalta R. and Drissi Y. (2002), "A Perspective View and Survey of Meta-Learning", Journal of Artificial Intelligence Review, 18 (2), pp.77-95.

[2] Saso D., and Bernard Z (2004), "Is Combining Classifiers with Stacking Better than Selecting the Best One?", Machine Learning, 54, Kluwer Academic Publishers, Netherlands, pp.255–273.

[3] Domingos Pedro (1998), "Knowledge Discovery via Multiple Models", Intelligent Data Analysis, 2, pp.187-202.

[4] Ting, K. M., and Witten, I. H. (1999), "Issues in stacked generalization", Journal of Artificial Intelligence Research, 10, pp.271–289.

[5] Breiman L. (1996), "Bagging predictors", Machine Learning, vol. 24, pp.123–140.

[6] Oza N. C. and Tumer K. (2008), "Classifier ensembles: Select real-world applications," Information Fusion, vol. 9, no.1, pp. 4–20.

[7] Dietterich, T. (2000), "Ensemble methods in machine learning", In Kittler, J., &Roli, F. (Eds.), First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, Springer-Verlag, pp. 1–15.

[8] Polikar R. (2006), "Ensemble based systems in decision making," IEEE Circuits System Mag., vol. 6, no. 3, pp. 21–45.

[9] Rokach L. (2010), "Ensemble-based classifiers," Artificial Intelligence Review, vol.33, pp.1-39.

[10] Liu M., Zhang D., Yap P. T., and S. D. (2012), "Hierarchical ensemble of multi-level classifiers for diagnosis of Alzheimer's disease", In proc. of Machine Learning in Medical Imaging (MLMI 2012), Lecture Notes in Computer Science, vol. 7588, pp. 27–35.

[11] Islam R. and Abawajy J. (2013), "A multi-tier phishing detection and filtering approach", Journal of Network and Computer Applications, vol. 36, pp.324–335.

[12] Xiao Jin, Xie Ling, He Changzheng, Jiang Xiaoyi (2012), "Dynamic classifier ensemble model for customer classification with imbalanced class distribution", Expert Systems with Applications, Volume-39, Issue 3, Pp. 3668-3675.

[13] Kelarev A.V., Stranieri A., Yearwood J.L., Abawajy J., Jelinek H.F. (2012), "Improving Classifications for Cardiac Autonomic Neuropathy Using Multi-level Ensemble Classifiers and Feature Selection Based on Random Forest", In Proceedings of the Tenth Australasian Data Mining Conference (AusDM 2012), Sydney, Australia, pp.93-101.

[14] Fumera, G. and Roli, F. (2005), "A theoretical and experimental analysis of linear combiners for multiple classifier systems", IEEE Transactions on Pattern Analysis and Machine Intelligence 27(6), pp.942–956.

[15] Kotsiantis SB. (2007), "Supervised machine learning: A review of classification techniques", Informatica, no.31, pp. 249-68.

[16] Melville P. and Mooney R. J. (2005), "Creating diversity in ensembles using artificial data", Information Fusion, vol.6, pp.99-111.

[17] Domeniconi, C. and Al-Razgan, M. (2009), "Weighted cluster ensembles: Methods and analysis", ACM Transactions on Knowledge Discovery from Data, 2(4), Article 17.

[18] Freund Y., Schapire R. (1996), "Experiments with a new boosting algorithm", Proceedings of 13th International Conference of Machince Learning, pp. 148-56.

[19] Koliopoulos A.K., Yiapanis P., Tekiner F., Nenadic G., Keane J. (2015), "A Parallel Distributed Weka Framework for Big Data Mining using Spark", IEEE International Congress on Big Data, IEE Computer Society, pp.9-16.