Using multiclassifiers in a case study with sugarcane land evaluation data

Saddys Segrera, and María N. Moreno

University of Salamanca, Department of Computer Science and Automatic, Plaza de la Merced s/n, 37008 Salamanca, Spain
{saddys, mmg}@usal.es
http://www.usal.es

Abstract. The combination of experts is an effective technique when the individual classifiers that are merged are precise and diverse. This paper presents an application of multiclassifiers in a real problem to predict the land aptitude of sugarcane areas. There are different combination strategies and multiclassifiers methods. We used Bagging and Boosting as ensemble methods with only one learning technique to achieve the classification process. Later, a hybrid method such as Stacking was used. These three multiclassifier methods yielded better results than the simple classifiers. Decision trees, nearest neighbor and Naive Bayes learning were the learning techniques applied.

1 Introduction

This paper presents a study about multiclassifiers, which currently have a wide apogee in the scientific environment, in contrast to the traditional classification techniques. Multiclassifiers take into account all valid hypotheses (one hypothesis per learner is combined). It means none of the consistent hypotheses is discarded; therefore the combination of the set predictions is carried out. The use of multiclassifiers has increased as a result of the overfitting problems they can solve, which makes possible to obtain better results with few training data. Moreover, they are able to decompose a complex problem in multiple sub-problems easier to understand, and additionally, they eliminate the not correlated errors of the individual classifiers [14], [19].

Multiclassifiers consist of a combination of different classifiers. Their function is to fuse individual classifier predictions in order to get the combination of all the previous predictions as a final result.

Multiclassifier methods may be heterogeneous. They can be divided into two main groups: one of them refers to the creation of a model from the combination of classifiers that use the same learning technique; the second kind of methods named hybrids consists of a combination of classifiers but with different learning techniques.

There are well known ensemble methods: Bagging [3], Boosting [8], Cross Validated Committees [17] (manipulate training examples); Random Subspace Method (manipulates input features) [12]; Error Correcting Output Coding (manipulates output targets) [7] and Randomization (injects randomness) [5].

Bagging, Boosting and Random Subspace Method were evaluated in [6] concluding that Boosting was the most precise among the three methods in problems without noise. The obtained results in [2] showed that Bagging was the best method in problems with noise, and Boosting presented the worst behavior.

Stacking [24] and Cascading [9] are hybrid methods used in many researches. Both methods differ in their architecture, their goals, and the internal use of cross validation, among others. In terms of architecture, Stacking has a parallel one, while Cascading shows a sequential architecture. On the other hand, the ultimate goal of Stacking is combining predictions; the goal of Cascading is to obtain a model that can use terms in the representation language of lower level classifiers. According to the third characteristic, Cascading does not use internal cross validation in contrast to Stacking that uses cross validation to generate a training set for learning the meta-level classifier.

There are different combination strategies to merge the output of individual classifiers. The abstract-level methods such as the majority vote [2], weight majority vote [15], behavior-knowledge space [13], and belief functions [25] represent one of these types of combination. Secondly, there are rank-level methods, as Borda count method [5] and weighted Borda count [22]. Another combination strategy is the measurement level fusion, which includes, for example: the simple average, the product, the maximum, the minimum, other statistic operators and weighted average [11].

Abstract-level methods can be applied to any ensemble of classifiers. However, the trained rules impose heavy demands on the quality and size of data set. Rank-level methods are suitable in problems with many classes. They can also be applied to soft outputs to avoid lack of consistency when using different classifiers and to simplify the combiner design; regarding they are not supported by theoretical underpinning and their results depend on the scale of numbers assigned to the choices. Finally, measurement level fusion, combines rules that can exploit a higher amount of information with respect to other results. In addition, complex combiners can be designed to exhibit classifiers with different performance and complex correlations. A normalization of the classifier soft outputs is required when different classifiers are being used. This is seen as a disadvantage as well as it is the use of large and very good quality datasets [18].

The classifier combination shows higher precision than any individual classifier in the set. This condition will be reached only if the individual classifiers are precise and diverse. That is, a classifier is considered precise when its error is lower than 0.5. Two or more individual classifiers are diverse when their output errors are not correlated.

The area of machine learning algorithms that deal with multi-agent systems is known as ensemble learning. Ensemble learning is based on the idea of having a set of weak learners (they can also be denoted as agents), which build together a strong learner through agreement mechanisms [10].

An ensemble of agents solves problems in the following way: each individual agent works out a part of the problem and makes its own prediction, and then, all those predictions are merged into a global decision [16]. Applications of multi-agents as multiclassifiers have been explored in [1], [21].

The objective of this paper is to use land evaluation data from sugarcane areas to predict the aptitude of each agricultural field by different individual classifiers and

multiclassifiers. Moreover, we want to demonstrate that the precision is increased with the use of a classifier fusion.

This paper is organized as follows: data analysis is given in section 2, evaluation and results from the use of different multiclassifiers are described in section 3. Finally, we summarize our application in section 4.

2 Data Analysis

During the development of this study the steps of the data mining process were followed as [4]. The objective is to apply data mining techniques by classifiers in data from the land physical aptitude categories dedicated to sugarcane in Cuba. These techniques will allow predicting the land aptitude categories from soil variables, climate and agricultural factors. Once precise individual classifiers are built, it will be proven how the use of multiclassifiers increases the precision.

The data source of this study is the National Sugarcane Research Institute of Cuba. The database has 1000 registers that correspond to sugarcane crops and twelve variables (soil slope, stones, rocks, salinity, soil pH, cation exchange capacity, drainage, compaction, rains, soil effective depth, agricultural yield cluster and land aptitude category). There are two numerical attributes and ten nominal attributes. The attribute value that will be predicted in this paper is represented in the label "eval", belonging to the land physical aptitude for sugarcane areas.

The Mineset software [20] was used in order to assure the analysis of data quality during pre-processing step. This tool was selected because its visualization capacity is more detailed and more illustrative than WEKA (Waikato Environment for Knowledge Analysis) 3.4 from University of Waikato [23]. WEKA was also used to execute algorithms because its variety of learning techniques and algorithms is larger than Mineset.

In Mineset, the option Statistics Viewer was used to visualize if there are variables with values significantly distant of the set. The visualization of the value distribution was obtained by histograms, in the case of the nominal attributes, and boxplots for the numerical attributes (Fig. 1). The graphics show that the variables do not present outliers.

Only in the case of the numerical variable "prfu" corresponding to the soil effective depth, the mean value of this variable is distant of the maximum value. From a total value of 1000, 78 of them are abnormally high; although, it is necessary to mention, this is not caused by human errors. The soils dedicated to sugarcane crops in Cuba usually, are not very deep, but there are some areas where it is possible to find high values of soil effective depth (very deep) and others with very small ones. All the attribute values belong to the established real data for the areas that integrate the data domain. It means that registration errors have not been made. In spite of attributes such as "ph", "cic", "sali" and "agrup" are not equally distributed among the different categories, it does not indicate that the categories with a few number of values are mistakes.

A pie chart (Fig. 2) was made to observe the behavior of the variable "eval", which is the label of this problem. As it can be seen in the chart, land aptitude data are

well distributed among the four possible categories. This variable is represented by the values “A1” for the extremely suitable areas; “A2” for those moderately suitable, “A3” for marginally suitable and “N” corresponding to those not suitable for sugarcane.

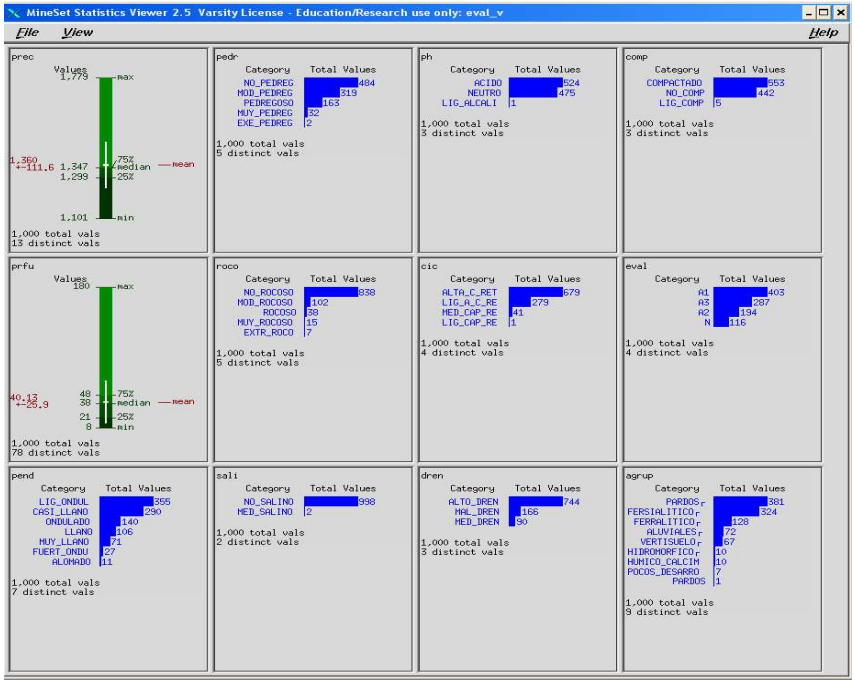

Fig. 1. Statistical behavior of variables by Mineset.

According to the results of the statistical analyses made to the data set, it is possible to affirm the data have high quality and they are not erroneous. That is, the data are reliable, there are not missed data and they follow an expected behavior. Data set has been delivered and validated by a distinguished specializing entity in Cuba. In general, this means the data do not have noise.

Several individual classification algorithms were tested using the tool WEKA to select those that present high precision in the prediction of the land physical aptitude category. Later on, multiclassifiers are created. Then, the precision increment with the use of multiclassifiers can be checked, in relation to the results using individual classifiers.

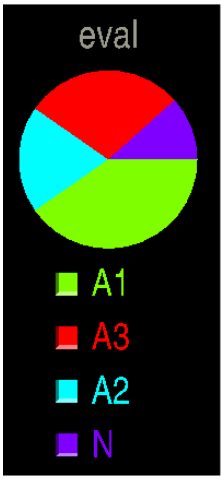

Fig. 2. Pie chart of the label "eval".

3 Evaluation and Results

Three simple methods were used: decision trees, Naive Bayes and learning based on the nearest neighbor. In all the experiments, the 10-fold cross-validation was used as test mode 15 times modifying the random number seed to be used. Table 1 shows the average precision and computational cost values of the three mentioned methods.

Table 1. Average precision (%) and time execution (sec.) values for three simple learning methods.

Algorithm	Precision	Time execution
Decision tree	90.17	2
Nearest neighbor	83.24	5
Naive Bayes	70.07	1

The results indicate that the best method for this dataset is the decision tree, followed by the nearest neighbor learning and, in third place, the Naive Bayes. Although Naive Bayes has less computational cost than decision tree, only a difference of one second, but the last one reaches an average error of 9.83 % while NaiveBayes has an average error value of 29.93 %. Nearest neighbor was the slowest among the three classifiers in the execution. Next, multiclassifiers were built using the same learning techniques: Bagging and Adaboost (Boosting) with the decision trees (it is the inductor with the highest individual precision obtained). In this manner we demonstrate the precision is increased in relation to the use of this learning technique in an individual way.

Table 2 illustrates the precision values obtained by the two multiclassifiers with decision trees. Both methods allow increasing the precision. Boosting was better than Bagging for this study according the precision achieved. Boosting using decision trees

registers the lowest average error value (6.50 %) but it is 6 times slower than the simple decision tree algorithm. In spite of multiclassifiers improve the precision they intensify the computational cost.

Table 2. Average precision (%) and time execution (sec.) values with decision trees for the individual classifier, Bagging and Boosting.

Algorithm	Precision	Time execution
Decision tree	90.17	2
Bagging	91.49	9
Boosting	93.50	12

Fig. 3 describes the precision behavior in relation to the number of iterations for Bagging and Boosting, using decision trees. The precision raise in both multiclassifiers is not constant according to the number of iterations. The highest variation occurs between 0 and 20 iterations and the highest values are reached with 40.

With the development of this study the results obtained confirmed the ones achieved in [25], because with this data set without noise, Boosting reaches a better precision than Bagging using decision trees.

When implementing Bagging and Boosting with the Naive Bayes algorithm, which is the learning technique that produced the worst precision in the individual way, the results were similar. Boosting surpassed Bagging (71.30 % and 71.00 %, respectively).

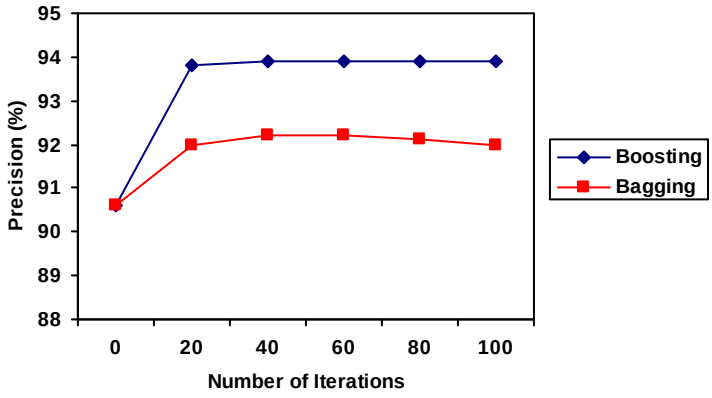

Fig. 3. Precision obtained with Bagging and Boosting with decision trees in relation to the number of iterations.

WEKA provides Stacking as a hybrid method. In the implementation of this method, the three initial algorithms (decision trees, Naive Bayes and learning based on the nearest neighbor) were the learning techniques used as base classifiers to build the model. Naive Bayes was also selected to learn the meta-model. The results reflect

Stacking increments the precision in comparison to the classifiers individually, but it was the slowest among the experiments made, see Table 3.

Table 3. Average precision (%) and time execution (sec.) values for three simple learning methods and Stacking.

Algorithm	Precision	Time execution
Decision tree	90.17	2
Nearest neighbor	83.24	5
Naive Bayes	70.07	1
Stacking	90.47	42

One of the National Sugarcane Research Institute of Cuba goals is the development of a Sugarcane Spatial Decision Support System (SSDSS). It will include the use of different classification techniques in data mining by multiclassifiers. The judgments (predictions) of human experts can be combined, and then it will take part of the system. This study constitutes an initial exercise that can be improved in order to incorporate classification models to the SSDSS.

The introduction of computer-advanced technologies will allow the integration of the sugarcane scientific advances to provide solutions to production problems. Therefore, it will permit that better results be reached in different research divisions as agronomical management, genetic and breeding improvement, sanitary control and agricultural services (fertilizers, varieties, pests and diseases, weeds, and others). All these elements can be joined in order to reinforce the current computer systems installed at National Sugarcane Research Institute of Cuba. Their integration based on multiclassifiers can fuse the knowledge of human experts (in the case of Cuba they are very specific and extensive). Making decision process can consider all criteria about sugarcane crops.

4 Conclusions

The appropriate combination of two or more classifiers can provide a more robust, reliable and efficient prediction than the individual use of classifiers. The combination can be implemented by means of intelligent agents representing weak learners, which together build a strong learner through agreement mechanisms.

There are differences in the combination forms and the final results of the multiclassifiers depending on the algorithms and the learning techniques used. The application of multiple classifiers introduces diverse classification approaches that offer higher flexibility in the final decision.

The hypothesis combination in multiclassifiers is an example of the most general and fundamental problem in the information integration from multiple sources. The main implication in the multiclassifiers is, in general, that they augment the precision in comparison with a single classifier. In addition, they reduce the over fitting, avoiding the selection of an extensive model.

The nature of the learning algorithms influences the classification precision with the same data set. Three different learning techniques were used in this research; the

individual classifiers were precise and diverse. Ensemble methods improved the precision value obtained by individual classifiers.

Boosting obtained better results than Bagging by using a dataset without noise.

A hybrid method (Stacking) was used. This method merged three different learning techniques and increased the precision value.

The use of multiclassifiers must be justified with a very significant precision gain in relation to the individual classifiers, due to; in general, the multiclassifiers take more time in their execution.

This experimental study allows a first exercise that can be included in the development of classification models using multiclassifiers for a Sugarcane Spatial Decision Support System.

References

1. Abreu, M.C.C., Canuto, A.M.P., Santana, L.E.A.S.: A comparative analysis of negotiation methods for a multi-neural agent system. Proceedings of the Fifth International Conference on Hybrid Intelligent Systems, November (2005) 451-456
2. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting and variants. Machine Learning, Kluwer Academic Publishers, Boston, Manufactured in The Netherlands, vol. 36, 1/2, (1999) 105-139
3. Breiman, L.: Bagging predictors. Machine Learning, Kluwer Academic Publishers, Boston, Manufactured in The Netherlands, vol. 24, 2, (1996) 123-140
4. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A.: Discovering Data Mining. From Concept to Implementation, Prentice Hall (1998)
5. De Borda, J.C.: Memoire sur les Elections au Scrutin. Historie de l'Academie Royale des Sciences, Paris (1781)
6. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning, Kluwer Academic Publishers, Boston, Manufactured in The Netherlands, vol. 40, 2, (2000) 139-157
7. Dietterich, T.G., Bakiri, G.: Solving multi-class learning problems via error-correcting output codes. Journal of Artificial Intelligence Research, vol. 2, (1995) 263-286
8. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. Proceedings of the 13th International Conference on Machine Learning, (1996) 148-156
9. Gama, J., Brazdil, P.: Cascade Generalization. Machine Learning, Kluwer Academic Publishers, Boston, Manufactured in The Netherlands, vol. 41, 3, (2000) 315-343
10. Gómez, F. E.: Automatized Multi-Agent System Design for Decision Making Problems. Tesis Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas, Puebla, Diciembre, (2005)
11. Hernández-Orallo, J., Ramírez, M.J., Ferri, C.: Introducción a la minería de datos. Pearson Educación, S.A., Madrid (2004)
12. Ho, T.K.: The random subspace method for constructing decision forest. IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 20, 8,(1998) 832-844
13. Huang, Y.S., Suen, C.Y.: A method for combining multiple experts for the recognition of unconstrained handwitten numerals. IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 17, (1995) 90-93
14. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, 3, (1998) 226-239

15. Littlestone, N., Warmuth, M.K.: The weighted majority algorithm. Information and Computation, vol. 108, (1994) 212-261
16. Ontañón, S.: Ensemble Cased Based Learning for Multi-Agent Systems. Tesi Doctoral. Consell Superior d'Investigacions Científiques, Institut d'Investigació en Inteligència Artificial, Escola Tècnica Superior d'Enginyeria, Universitat Autònoma de Barcelona, Bellaterra, Abril (2005)
17. Parmanto, B., Munro, P.W., Doyle, H.R.: Improving committee diagnosis with resampling techniques. Advances in Neural Information Processing Systems, vol. 8, eds. Touretzky, D.S., Mozer, M.C. and Hesselmo, M., MIT Press, USA (1996) 882-888
18. Roli, F.: Fusion of Multiple Pattern Classifiers. 8th National Conference of the Italian Association on Artificial Intelligence, September, Pisa, Italy (2003)
19. Sharkey, A., Sharkey, N.: How to improve the reliability of artificial neural networks. Technical Report Cd-95-11, Department of Computer Science, University of Sheffield (1995)
20. Silicon Graphics Computer Systems: MineSet. Technical report, Silicon Graphics Computer Systems (1998)
21. Sylvester, J., Chawla, N.V.: Evolutionary Ensembles Combining Learning Agents using Genetic Algorithms. Proceedings of the AAAI Workshop on Multi-Agent Learning, California, July (2005)
22. Van Erp, M., Schomaker, L.: Variants of the Borda count method for combining ranked classifier hypotheses. Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, Amsterdam, September, (2000) 443-452
23. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco (2005)
24. Wolpert, D.H.: Stacked Generalization. Neural Networks, Pergamon Press, vol. 5, (1992), 241-259
25. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems Man and Cybernetics, vol. 22, 3, (1992) 418-435