Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Ensemble Classification Using Agents

Nimisha Peddakam, Sreevidya Susarla, Annepally Shivakesh Reddy
CSE Department, CBIT, Telangana,
India

Abstract- Multi agent and data mining have been widely used in building large and complex system. A heterogeneous multi agent system improves accuracy of large and complex data mining task. A working agent is a data mining agent which is used to classify the data stream, as mining with single classifier is not accurate, to achieve accuracy, ensemble classification is used in agents which uses multiple learning algorithms to obtain better predictive performance. The proposed system implements parallel and incremental mining algorithms, data can be divided into partitions that are processed by all the agents in the systems and the results from the partitions are then merged to improve the efficiency. The root mean square value and the confusion matrix of the highest accurate classified data stream are used to compare with those of the classified data stream using single classifier.

Keywords: Classification, Cross validation, Agents, Accuracy, Decision table, Single agent classifier, Multi agent classifier.

I. INTRODUCTION

Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late. What is all the excitement about? This article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. The article mentions particular real-world applications, specific data-mining techniques, challenges involved in real-world applications of knowledge discovery, and current and future research directions in the field. Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD). At an abstract level, the KDD field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact (for example, a short report), more abstract (for example, a descriptive approximation or model of the process that generated the data), or more useful (for example, a predictive model for estimating the value of future cases). At the core of the process is the application of specific data-mining methods for pattern discovery and extraction.

II. LITERATURE SURVEY

Ensemble classification uses multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble refers only to a concrete finite set of alternative models, but typically allows for much more flexible structure to exist between those alternatives.

III. OVERVIEW

Supervised Learning algorithms are commonly described as performing the task of searching through a hypothesis space to find a suitable hypothesis that will make good predictions with a particular problem. Even if the hypothesis space contains hypotheses that are very well-suited for a particular problem, it may be very difficult to find a good one. Ensembles combine multiple hypotheses to form a (hopefully) better hypothesis. In other words, an ensemble is a technique for combining many weak learners in an attempt to produce a strong learner. The term ensemble is usually reserved for methods that generate multiple hypotheses using the same base learner. The broader term of multiple classifier systems also covers hybridization of hypotheses that are not induced by the same base learner. Evaluating the prediction of an ensemble typically requires more computation than evaluating the prediction of a single model, so ensembles may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation. Fast algorithms such as decision trees are commonly used with ensembles (for example Random Forests), although slower algorithms can benefit from ensemble techniques as well.

IV. ENSEMBLE THEORY

An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built. Thus, ensembles can be shown to have more flexibility in the functions they can represent. This flexibility can, in theory, enable them to over-fit the training data more than a single model would, but in practice, some ensemble techniques (especially bagging) tend to reduce problems related to over-fitting of the training data. Empirically, ensembles tend to yield better results when there is a significant diversity among the models. Many ensemble methods, therefore, seek to promote diversity among the models they combine. Although perhaps non-intuitive, more random algorithms (like random decision trees) can be used to produce a stronger ensemble than very deliberate algorithms (like entropy-reducing decision trees). Using a variety of strong learning algorithms, however, has been shown to be more effective than using techniques that attempt to dumb-down the models in order to promote diversity.

V. SYSTEM ANALYSIS

A. Existing System

Existing work had the objective of working with single classifier with a single agent system which did not result accuracy. Single agent system with single classifier.

B. Proposed System

A heterogeneous multi agent system to improve large and complex data mining task. As classification using single agent system is not so accurate, to achieve accuracy, we propose a multi agent system with ensemble classification.The proposed system has following advantages

- It can be used to solve problems that are too large for centralized system.
- Agents are capable of independent actions on behalf of a user.
- Agents can learn from their own experiences.

A data set which is taken from the UCI repository is given as the input to the system. Then, depending upon the size of the data set, the agents are created by the coordinator. Each agent consists of three classification algorithms. The coordinator, then, partitions the data set and sends each partition to all the agents. The best classifier i.e. the most accurately classified data set is then selected. This is called cross validation. Then, the accurate data set is chosen after the subsequent classification. Similarly, the same data set is sent to a single-agent system. The data set is classified by only one agent here. Now, the comparison is shown between the accuracies of the classified data sets. A data stream is sent as the input. The agent manager partitions the data stream and sends each of the partitioned streams to all the mining agents. Each agent contains three distinct classifiers.Each algorithm in each one of the mining agents performs the classification with some accuracy.

VI. CROSS VALIDATION

It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice in a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset). The goal of cross validation is to define a dataset to "test" the model in the training phase. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set).

VII. ENSEMBLE THEORY

A data stream is sent as the input.The agent manager partitions the data stream and sends each of the partitioned streams to all the mining agents.Each agent contains three distinct classifiers. Each algorithm in each one of the mining agents performs the classification with some accuracy.

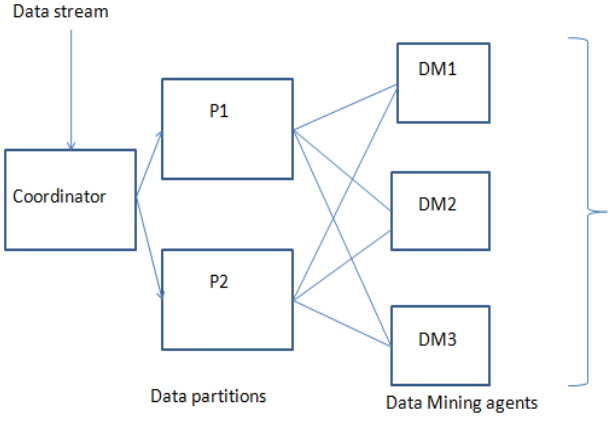

Fig. 1 Diagram depicting ensemble Classification Using three agents

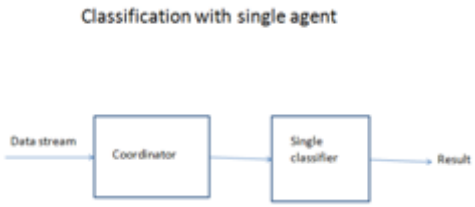

Fig 2 Classification Using Single Agent

VIII. MODULES AND FUNCTIONALITIES

A. Module 1- Partitioning

The dataset chosen by the user is taken in the ARFF format. Depending upon the size of the dataset, it is partitioned into some number of folds. In this system, the datasets with less than 1000 instances have been divided into 10 folds and those with instances more than 1000 and less than 5000, have been divided into 20 folds. Now, these folds have been divided into various partitions depending upon the size of the dataset. Each such partition is sent to a fixed number of agents. The number of agents to be employed for the task is also dependent on the size of the dataset. Here, the dataset with 1000 instances is divided into 10 folds and two partitions each containing 5 folds. Two agents are used to perform the classification. The dataset with 1000-5000 instances is divided into 20 folds and four partitions each containing 5 folds. Four agents are used for classifying these four partitions.

B. Module 2- Classification of partitions by JADE Agents

Each partition obtained, as described in the above module, is sent to all the agents considered for performing the classification. Each agent contains four different classifiers such as J48, Decision Table, etc. The agent classifies the partition sent to it by all the algorithms present in it. Then, by comparing all the thus obtained accuracies, the best accurate classifier's result and confusion matrix are considered for further processing. The other classifiers' accuracies are ignored.

C. Module3-Calculation of the final accuracy for the dataset

Each partition is sent to all the agents and from each agent the best accuracy is considered. For example, consider there are two partitions and two agents. The first partition is sent to both the agents. Then, the accuracies returned from the both agents are compared and the highest one is considered. Similar operation is performed for all the partitions. Now, we have the best accuracies of each partition. The final accuracy is calculated by computing the average of all the above obtained accuracies of all the partitions.

D. Module 4-Generation of accuracy table

The user is allowed to choose five different datasets. The above stated operations are performed for each dataset and the final accuracies are obtained by the usage of Ensemble classification using agents. A table is generated which contains the accuracies of the classified dataset by using J48 classifier, Decision Table classifier and Ensemble classifier. J48 and DecisionTable are pre-defined classifiers in Weka. It is observed from the table that the accuracy of ensemble classifier is more than the other two classifiers.

E. Module 5-Generation of graph

A bar graph is plotted for the table generated above. X-axis contains the five datasets and for each dataset, there are three bars. Each bar representing one classifier (J48, DecisionTable and Ensemble).Y-axis consists of the accuracies. In the graph, we can observe the change in the accuracies for the corresponding classifiers. Thus, the technique implemented by this system is more efficient than the existing systems.

IX. RESULTS AND DISCUSSIONS

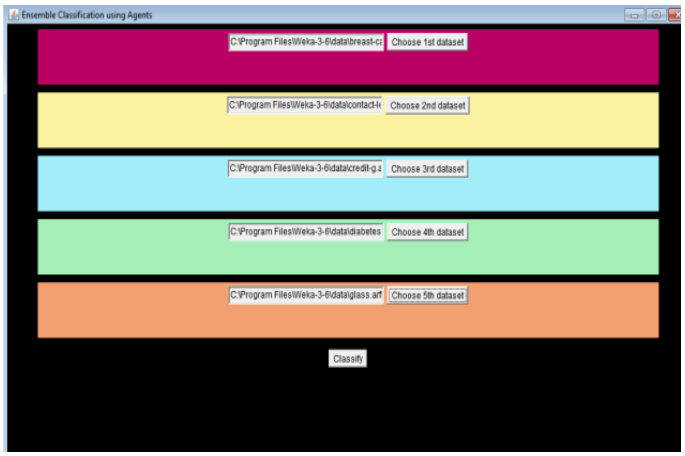

Fig 3 After selecting the data sets from UCI Repository, the user needs to click on the classify button to classify the data in the datasets

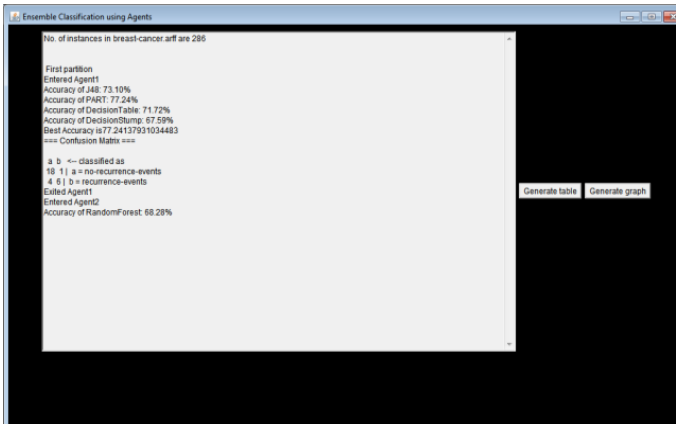

Fig 4 Output which consists of the each and every classifier's accuracy, the confusion matrix of every classifier, the accuracy of each mining agent

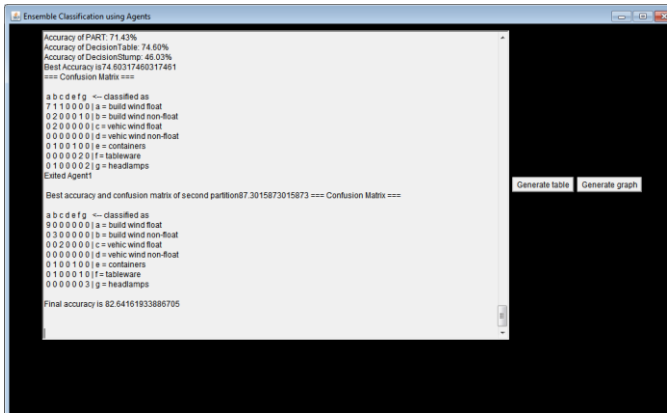

Fig 5 Displays the best maximum accuracy of every classifier in each mining agent along with the confusion matrix of the best classifier in every mining agent

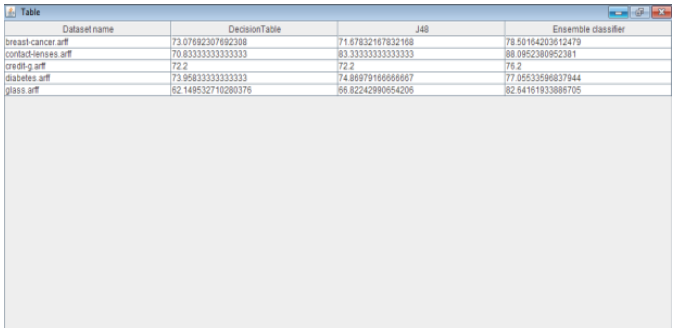

Fig 6 Displays a table comparing the accuracy which includes the general classifiers like J48 and Decision tree and the ensemble classification done in the project

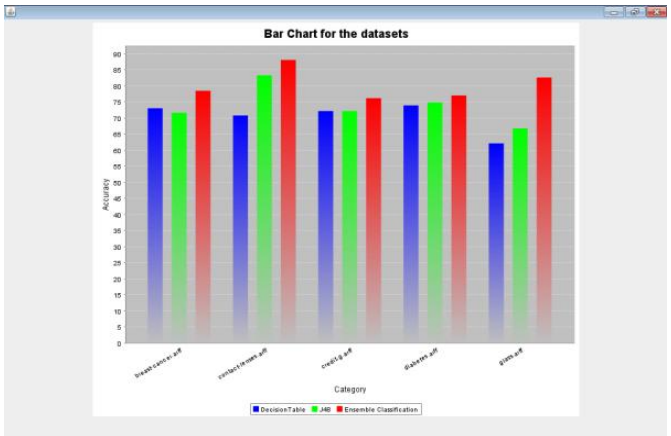

Fig 7 Displays the graph by comparing the accuracies of the classifiers on y axis and taking different data sets on the x axis

X. CONCLUSION

The main objective of this project is to implement parallel and incremental mining algorithms, data can be divided into partitions that are processed by all the agents in the systems and the results from the partitions are then merged to improve the efficiency. The root mean square value and the confusion matrix of the highest accurate classified data stream are used to compare with those of the classified data stream using single classifier. A graph has been plotted for multiple data sets by comparing multi agent system with single agent system.

REFERENCES

[1] Cheeseman, P. 1990. "On Finding the Most Probable Model. In Computational Models of Scientific Discovery and Theory Formation", eds. J. Shrager and P. Langley, 73–95. San Francisco, Calif.: Morgan Kaufmann.

[2] M. Berry, G. Linoff, *Data Mining Techniques: for Marketing, Sales and Customer Support*, Wiley

[3] U.S. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy,*Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.

