Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN AUTOMATED FRAMEWORK FOR AGENT BASED FEATURE SELECTION AND CLASSIFICATION

[1] B.KALPANA, [2] DR V. SARAVANAN, [3] DR K VIVEKANANDHAN

[1] *Ph.D* Research Scholar Bharathiar university &
A.P, Department Of Computer Science & Applications.,PSG College Of Arts And Science,Coimbatore,India
[2] Professor & Director,MCA Dept., Sri Venkateswara College Of Computer Applications And Management , India
[3] Professor,School Of Management ,Bharathiar University,Coimbatore,India
E-mail: [1] bkalpana.psg@gmail.com , [2] tvsaran@hotmail.com ,[3] vivekbsmed@gmail.com

ABSTRACT

The proposed framework for agent based feature selection and classification system works with the help of software agents. These agents are rule based and are able to guide the user in feature selection and classification. With number of feature selection and classification algorithms available the framework paves way for an integrated approach in feature selection and classification. Initial results with partially implemented system prove to be promising in the field of machine learning. The algorithm was used on a live web site data for three years and the results were mined using the framework. The results give hope to show that agent based decision making can be very useful for persons who do not have idea of mining but are decision makers.

Keywords: *Feature Selection, Classification, Agent Based, Mining*

1. INTRODUCTION

When faced with a new problem or situation, people study past experiences and reuse those experiences to take a decision [1][2]. Software agents are programs that work on behalf of user to take action [3]. Software agents can capture the experience of a user and propagate that to a novice user. This has been tried in web mining applications [4][5][6], robot control[7], travelling[8], agriculture[9] etc.

Data mining is the process of finding novel, useful patterns from available data.[10]. An automated data mining framework helps in aiding the user to automatically collect data from various sources, preprocess, select appropriate features and classify data. The software agent helps to store user navigation pattern (the selection of parameters, techniques) and help novice user to use the framework successfully. An agent can work as a collector, pre processor [11] and classifier [12], recommending assistant [13][14] to a user. Mining is accomplished easily only when the person knows all about mining algorithms and seed parameters. An automated framework can help user or decision makers when they have little or no knowledge about mining techniques and parameters for the algorithm chosen.

This paper proposes an automated framework and the use of software agents in assisting to select various features and classification algorithm to classify the data. The agent automatically stores the type of data and the algorithm used in storage for future reference. The following sections discuss the framework, feature selection and classification done on various data sets. This framework accepts and uses the skill of different users and stores results during each stage for future reference.

2. AGENT BASED FRAMEWORK

In the proposed agent based framework the software agents are used to assist a user who has no prior knowledge of mining techniques. Here the system has three agents which are used for

Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

recommendation, feature selection and classification. In the framework the role of recommendation agent is a vital one. The agent stores previous users mining information to guide a user in suggesting top queries given to the application. When the query is selected, appropriate dataset is passed to the feature selection agent .The role of feature selection

agent is to select relevant features from available dataset for classifying with maximum accuracy. The feature selection agent has a set of algorithms available which can be used for a particular data type in database. The results are passed to the classification agent. The classification agent uses a meta classifier such as stack generalization to do the classification. The data type again determines the classifying algorithms selected by the agent. The results of classification are stored in database to be used by recommendation agent at a later time.This framework has a assigned work for each agent described.

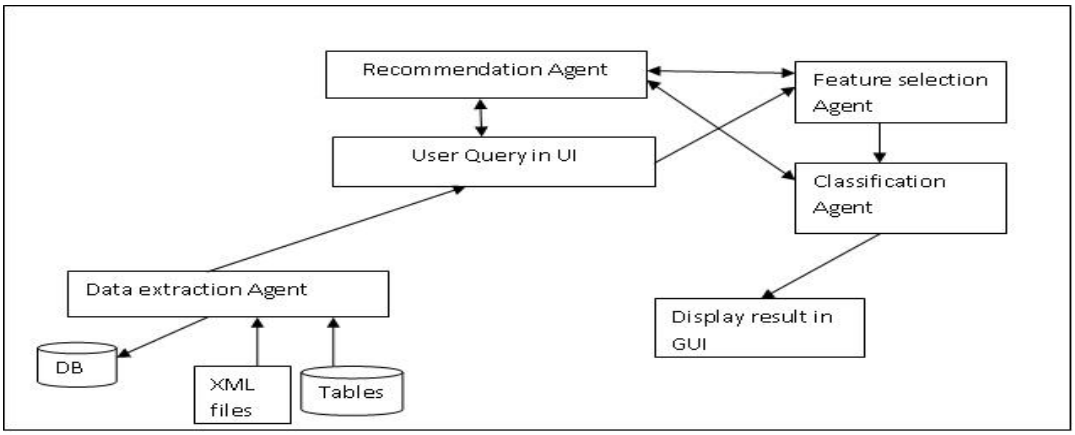

Fig 1: Automated Framework

A. Data Extraction Agents

Many framework for agents have been proposed [15][16] for intelligent business mining. In any given enterprise the type of data to be mined is known beforehand be it medical data, sales data, Employee data etc. In the proposed framework the agents collect data from different sources. For example most of the information is available online in XML format. In a distributed environment the data is available from different sources and formats like tables, HTML, emails.

Fig 2: Framework Developed In Java For Combining Multiple XML Files

B. Feature selection agents.

The software agents select the features required for classification based on user query. Preprocessing is done based on the query. There is no intelligent method to select a required algorithm even though the number of available algorithms has increased. An end user is not only required to know the domain well but also is expected understand technical details of available algorithms in order to make a "right" choice[17].Therefore, the more algorithms available, the more challenging it is to choose a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

suitable one for an application. Consequently, a big number of algorithms are not even attempted in practice and only a couple of algorithms are always used. Therefore, there is a pressing need for intelligent feature selection that can automatically recommend the most suitable algorithm among many for a given application.

Feature selection is a process that selects a subset of original features. An evaluation criterion is used measure the optimality of a feature subset .As the dimension of the domain expands the number of features N, will increase. Many problems related to feature selection have been shown to be NP-hard [18]. A typical feature selection process consists of four basic steps namely, subset generation, subset evaluation, stopping criterion, and result. Knowledge about mining the dataset and information regarding the data are two key determining factors for feature selection. The knowledge factor covers purpose of feature selection, expected Output Type, and S=N Ratio—the ratio between the expected number of selected features S and the total number of original features N and Time concern. The data factor covers Class Information, Quality of data, and N=I Ratio—the ratio between the number of features N and the number of instances I and Feature Type. Given four different algorithms for feature selection, then selection can be done based on the dataset and time efficiency or accuracy of predictions of the algorithms. Algorithm1 helps to select the required feature selection algorithm based on classification accuracy. Even though initially all algorithms would run on the dataset,the results of the dataset will be stored permanently in database. Given n number of feature selection algorithms, the job of agents is to choose an appropriate algorithm based on the type of attributes, number of attributes, classes and algorithms available.

Algorithm

Agent_FS(String[][] dataset)

// Given N features, we have to select relevant M features based on M/N ratio.

// Flag is set to 1 if same features classified previously as determined by Same_dataset() method

// A set of algorithms is available and loaded for each type of data to FS[] array from db

Begin

Step 1. If required Preprocess the dataset //like cleaning and removing noise

Step 2. If (flag= Same_dataset(dataset)) goto step 6

Step 3. Run FS[1] and get M_{best} and CP_{best} // M and CP represent no of relevant attributes and //classification accuracy

Step 4. For i=2 to n do

a. Run FS[i] on dataset and obtain M' and CP

b. If M' < M_{best} && M'> 1 then
If abs(CP_{best} – CP) <= Θ //threshold level by user
Begin
M_{best} = M'
CP_{best} = CP
end

endfor

Step 5. Store the result in DB for future use and goto step 7

Step 6. For the same dataset accept relevant features from recommendation agent along with M_{best} and CP_{best}

Step 7. Return results to classification agent.

End.

The agent can also suggest what result the previous selection of query prompted into selection of this algorithm. The knowledge factor discussed previously comes in handy here. The purpose of feature selection is previously known in this algorithm.

C. Classification Agent

The role of the classification agent is to accept input from feature selection agent and build a meta classifier using stack generalization as introduced by Wolpert [19]. The set of base classifiers are stored in database for each data type. As mentioned in [20] the agent loads the classifiers according to data type of the current query . A meta classifier is added to the stack which classifies the dataset based on the resut of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

base classifiers. The data set is split into two as training and test data. The training data is classified using base level classifiers. The predictions are performed and the results of the classification are stored for the meta level classifier to perform on the test data set.

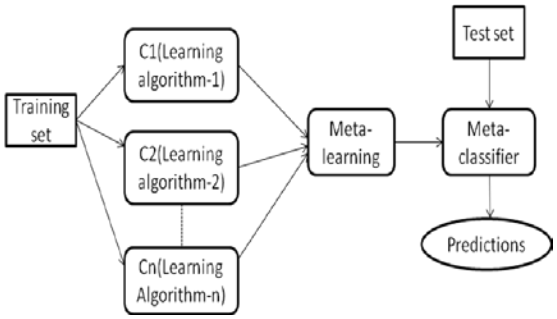

Fig 2. Learning Of Meta Classifier

.

The base classifier gives the classification with is stored along with class probability function and the confidence of the predictions. The confidence level helps the meta classifier to predict the correct class for the selected instance. The role of the classification agent is to split the dataset for training and test as the parameter set by user or predetermined minimum. The agent has to choose the appropriate classifiers in the database and load it for base and meta level.

D. RECOMMENDATION AGENT

The role of the recommendation agent is to store the selection of various items in the GUI and present it to user for easy navigation. The user may have an idea of mining algorithms or can use the suggestions by recommendation agent. The recommendation agent stores user navigation pattern along with selected items for future reference. For example if the state space consists of all possible combination of navigation, then the job of the agent is to group all similar interest

Table 1. Datasets Taken For Experiment

Dataset	Instances	Attributes	No. of classes
Iris	150	5	3
Soybean	683	63	19
Segment	1500	20	7

users and give them directions for recommendation. To develop mining tasks for navigation pattern, we need to estimate how similar two mining patterns are. We introduce a metric that estimates the similarity based on the total number of common categories that coexist to the total number of distinct categories. According to [21] the sequential subsequence can be obtained from decisive users and help the indecisive users. For example, if association between an item-X with that of n number of other item. We can use this metric to identify users who share common navigation behaviour and search interests. These are people who search for new information in a similar way. They form the "decisive group". We can use k-means algorithm to group similar interest users and the agent help the indecisive users with the clusters obtained to recommend popular predictions.

3. Experiments And Results

A. UCI Dataset

The framework was implemented using Java. Initially only the work of feature selection agent and classification agent were conducted. Three datasets from UCI repository were taken for consideration as shown in Table 1. The data types chosen were nominal. A set of classifiers like Naïve Bayes , Decision tree and K-nearest neighbor were stored in Database for nominal type. The result of classification run without feature selection is shown in Table 2.

Table 2. Classification Done Without Feature Selection

Dataset	Naive Bayes	C4.5	KNN	METACLASSIFIER
Iris	95.53	94.73	95.73	95.33
Soybean	91.36	88.74	84.31	90.50
Segment	80.17	96.79	95.25	95.69

Table 3: Classification Done Using Feature Selection And Meta Classifier

Dataset	Naive Bayes	C4.5	KNN	MC--C4.5
Iris(2,cfs)	96.08	96.08	96.08	96.08
Soyabean(13,IG)	.99	94.1	95.94	98.19
Segment(7,Cfs)	83.53	95.88	92.16	96.86

The feature selection agent had correlation based feature selection and IG as its feature selection algorithms. The result of feature selection agent was passed to classification agent and the results are as shown in Table 3. The result of classification shows a marked accuracy by using this method. The graph in fig 4. show the mining results using feature selected meta

Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

classifier for iris dataset. Fig 4 shows a steep increase in accuracy for soyabean dataset. The framework was developed using Java and used library imports for classification algorithms.

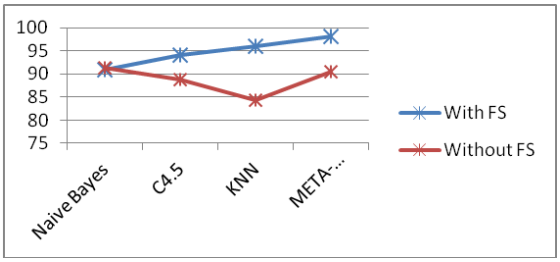

Fig 3: Classification Accuracy Achieved By The Framework

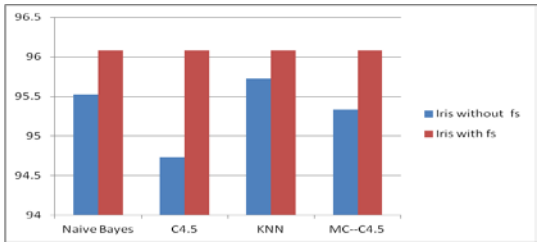

Fig :4 Accuracy Improvement With FS For IRIS Dataset

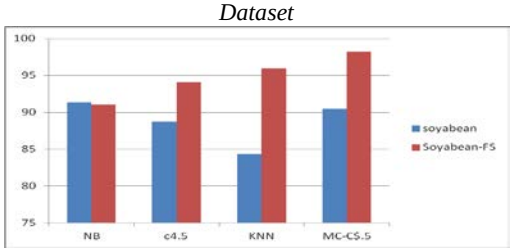

Fig 5. Soyabean Result Accuracy Improvemet

B. Website Data

Mining was done on SQL dump provided by online bus travel ticket booking website ticketgoose [22] . The results of mining for passengers travelled, how they came to know about the website, the most sought after destination were classified and shown fig 6,7 and 8 respectively.

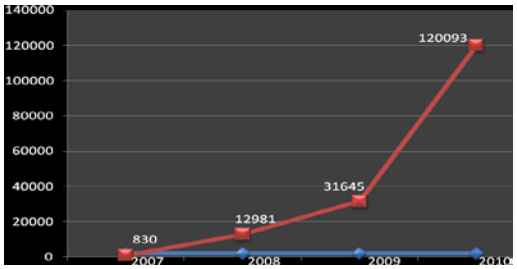

Fig 6. No Of Persons Travelled Between 2007 And 2010

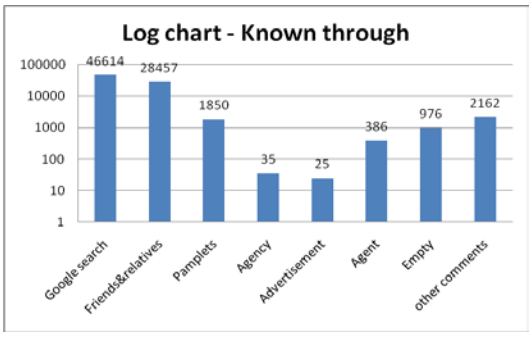

Fig 7. How Users Came To Know About Website

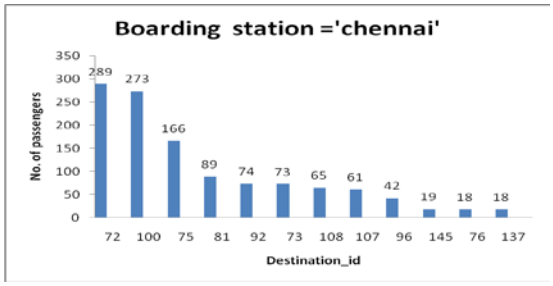

Fig 8. No. Of Users Boarding In Chennai Along With Destination Id.

Since the data was not preprocessed for empty values Fig. 7 shows a classification result as empty. In Fig 8 the source and destination places are marked using station-ids. The figure shows different destination Ids for travelers boarding in Chennai.

4. DISCUSSIONS

The results of the imparted framework show an increase in mining accuracy. If a system is developed with a set of known classifiers for the data to be stored then the person who makes decisions based on data stored requires no knowledge of mining [23]. The number of base classifiers in our experiment has been limited to three in our experiment. As concluded by [24] the result of the meta classifier is not affected by the number of base classifiers. The type of data is only of nominal type taken in our experiment. It has been proved that large number of attributes will show a greater accuracy with meta classifier [25]. The classification accuracy achieved on different datasets with different dimension reduction strategies is sensitive to the type of data. Our result for soyabean dataset shows the increased accuracy. Soyabean has IG as its selected strategy. From the original 63 attributes it selects only 13 attributes and gives greater accuracy. The other two datasets iris and segment has less attributes but more instances thus

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

enabling cfs strategy to act as feature selection algorithm. As for the website data it was cleaned, preprocessed and stored in a table for classification. The original data consisted of 63 tables and the table for booking a bus ticket had approximately around one lakh rows. The data obtained was preprocessed for noise reduction. The results obtained helped the decision makers to suggest increase in certain bus routes. The number of passengers travelling in buses increase during festival time on certain routes. This was captured during our analysis. The result in mining using stacked classifier with feature selection shows better result than using a single classifier. The reduction algorithm in an application context is still an active research area. The classification accuracy achieved using reduced data set is preferable than full data set.

5. CONCLUSION

We developed a framework where different feature selection algorithms and classification algorithms for nominal data were available for a single data set. We have studied various feature selection algorithms and their impact on different classification algorithms. Our experiments show the importance of dimensionality reduction and which in turn results in higher classification accuracy. Users who are not aware of different mining algorithms can use a stacked meta classifier with predetermined parameters for classification. The role of the recommendation agent to store the results of previous classification and display required results. It can also suggests various feature selection algorithms for different types of data. The initial framework was static one. With the results showing a promising trend a methodology for storing previous results and preferences can be implemented in future.

ACKNOWLEDGMENT

We thank Arun Athiappan of Ticket goose for providing SQL Dump of Ticketgoose data of 2007 to 2010.

REFERENCES:

[1] Y.Goa, I. Zeid and T.Bardez “Characteristics of an effective design plan to support re-use in case-based mechanical design”. Knowledge based systems, 10:337-50, 1998.

[2] D Khadilkar L.Stauffer “An experimental evaluation of design information reuse during conceptual design ,Journal of engineering design 7(4):331-9,1996.

[3] B. Mobasher, R. Cooley, J. Srivastava. “Automatic Personalization Based on Web Usage Mining”, Communications of the ACM, Volume 43, Number 8 2000.

[4] J.D. Vel´asquez, A. Bassi, H. Yasuda, and T. Aoki “Mining web data to create online navigation recommendations,” Proc. 4th IEEE Int. Conf. on Data Mining, pp.551–554, Brighton, UK, Nov. 2004.

[5] J. Ji, Z. Sha, C. Liu, and N. Zhong, “Online recommendation based on customer shopping model in e-commerce,” Proc. IEEE/WIC Int.Conf. on Web Intelligence, pp.68–74, Halifax, Canada, Oct. 2003.

[6] X. Fu, J. Budzik, and K. J. Hammond,“Mining navigation history for recommendation,” in Proceedings of the International Conference on Intelligent User Interfaces (IUI '00), pp. 106–112, 2000.

[7] Posadas, J.L., Poza, J.L., Simo´ , J., Benet, G., Blanes, F., “Agent-based distributed architecture for mobile robot control” Engineering Applications of Artificial Intelligence 21 (6), 805–823, 2008.

[8] Schiaffino, S. & Amandi, A. “Building an Expert Travel Agent as a Software Agent” Expert System with Applications,2008

[9] Ye-Ping, Z., Shi-juan, L. & Yue, “Z.Application of the Agent in Agricultural Expert System Inspection Software” Agricultural Science in China, 7(1).117-12,2008

[10] P-N, Steinbach M, Kumar V Introduction to data mining, 2006 Pearson Addison-Wesley.

[11] Othman, Z.A.,Abu Bakar, A.; Hamdan, A.R.; Omar, K.; Shuib, N.L.M.” Agent based preprocessing”
International Conference on Intelligent and Advanced Systems, IEEE Xplore, pg 219 - 223 , 2007.

[12] Bakar, Azuraliza Abu Othman, Zulaiha Ali Hamdan, Abdul Razak Yusof, Rozianiwati Ismai “Agent based data classification approach for data mining” International Symposium on Information Technology, IEEE Explore pg 1 - 6 2008,

[13] Balabanovic, M. and Y. Shoham, “Fab: Content-based Collaborative Filtering

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Recommendation," Communications of the ACM, 40(3): 66-72, March 1997.

[14] Pazzani,M. J., Billsus, D."Content-Based recommend -dation Systems". Lecture Notes in Computer Science, 2007, 4321, p. 325-341.

[15] Tung Bui, Jintae Lee, "An agent-based framework for decision support systems", Decision Support Systems, Volume 25, Issue 3, , Pages 225-237, April 1999

[16] Seydim A. Y.: Intelligent Agents: A Data Mining Perspective, Department of Computer Science and Engineering, Southern Methodist University, 1999

[17] Huan liu, Lei yu "Towards integrating feature selection algoritms for classification and clustering", IEEE transactions on knowledge and engineering, Vol 17, no.4, 2005

[18] A.L. Blum and R.L. Rivest, "Training a 3-Node Neural Networks is NP-Complete," Neural Networks, vol. 5, pp. 117-127, 1992.

[19] Wolpert, D.H ,"Stacked Generalization". *Neural Networks*, 5(2):241–260 1992.

[20] Todorovski.L , Dzeroski.S., "Combining Classifiers with Meta Decision Trees", Machine Learning Journal, Volume 50, pp 223-249,2003

[21] P-miner: Using user navigation patterns for personalizing topic directories, viewed on September 10, 2011 http://www.dblab.ntua.gr/~pbour/p-miner/

[22]http://www.ticketgoose.com

[23] B.Kalpana,V.Saravanan, K. Vivekanandhan ," A Framework For Mining Business Intelligence – A Boon To Non Mining Experts" IJRCM , Volume no. 2 (2012), Issue no. 2 (February)

[24] Džeroski S, Ženko B (2004) Is combining classifiers with stacking better than selecting the best one?. Mach Learn 54(3):255–273

[25] Lior Rokach ," Ensemble-based classifiers", Artificial Intellegence Review (2010) 33:1–39.