HYBRID CLASSIFICATION TECHNIQUES

Hanana K H[1], Maya Mohan[2], and Sruthy Manmadhan[3]
[1]M.Tech First Year Student, [2,3]Assistant Professor
Department of Computer Science and Engineering,
N.S.S College of Engineering, Palakkad
Email: [1]henashines@gmail.com, [2]mayajeevan@gmail.com, [3]sruthym.88@gmail.com

***Abstract*—Knowledge is the important part of human life. There are number of data mining systems that are available today and have many challenges in this field. Since the dataset that we are dealimg today is too big and it contains the data which shows correlation between different classes, classiication of such datas are complicated. Different types of classification techniques are available in data mining like Decision tree, C4.5, Bayesian networks, Neural networks, Support vector Machine, Association rule, K-NN, CART etc. Recently, there exist various soft computing techniques also like Genetic Algorithm, Ant Colony Optimization, Firefly Algorithm, Cuckoo Search, Artificial Bee Colony, Levy Flight etc. They are also combined with the various other techniques like rough set, fuzzy logic and neural network etc to obtain an effective syste to classify those objects. This paper presents an extensive review of literature on various hybrid technologies used for classifying those uncertain datas.**

Keywords: Data mining, Classification, Hybrid, Weights, Feature selection.

I. INTRODUCTION

Today, the intense lump in data due to the large scale mechanization and computerization of business, easily available and affordable hardware, software and data collection and management tools. The information is hidden in this large data. To handle this huge amount of data and to get the hidden information from the data is a very time consuming process. Also, data is stored in different forms. To get the effectual and fruitful results from these different forms is a very challenging task. Therefore, to extract this hidden information from the data, data mining process is used.

When the features are extracted these data need to be classified according to their class labels. Since the source of the data are different, classifying the according to our need is a difficult task. Usually many data fall on the boarder of different classes. So they shows the properties of many different classes thus making the classification cumbersome. They cannot be put into a singleton class. They need to be processed well to reduce missclassification errors.

Through this paper I am giving an overview of some hybrid classification techniques.

II. CLASSIFICATION IN DATAMINING

In data mining, Classification is the process of finding a model that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is not known. The model is based on the analysis of data objects whose class label is known. The derived model is represented in different forms like if-then rules, decision tree, mathematical formulae and neural networks. Classification process predicts the categorical labels. Classification process consists of a two-phase. In first phase, a model is constructed to explain a predetermined set of data classes. This is the training step. In second step, this classifier or model is used for classifying future or unknown objects. The classifier accuracy is the percentage up to which model correctly classified the test samples. Hybrid classification combines the strength of more than one classification model to enhance the performance of classification thus reducing misclassification errors.

III. DIFFERENT TYPES OF CLASSIFICATION TECHNIQUES

In data mining, there are different methods which are widely used for classification of data. They are:

1) Decision tree induction
2) Artificial Neural Network
3) Association rule analysis
4) Support Vector Machine
5) K-nearest neighbour
6) Bayesian Classification
7) Rule Based Classification

In data mining, there are some advanced methods which are also used for classification. These are:

1) Swarm Intelligence
2) Genetic Classifiers
3) Rough set approach
4) Fuzzy set approach

IV. RELATED WORKS

A. Hybrid Classification System for Uncertain Data

Most of the data we are dealing today contains many objects from many different classes. The proposed algorithm [1] uses KNN algorithm[2] to classify these data in three different approaches. Here SOM is used to obtain a lower dimensonal object space[3]. The first approach is for those objects which can be correctly classified to a singleton class. That is all the closest nodes to this object lies in the same class. So this object can be confidently classified into the class of those neighbouring nodes according to simple hard classification. Some objects have

neighboring nodes lying in different classes but the closest node is not labelled as boarder node. Such objects can be classified using fuzzy classification by calculating the support degree functions of each objects using a weighting factor. A basic belief assignment(BBA) is computed and then BBAs of different objects are combined. The combination result here contains only the singleton elements and the object will be placed in that label. Credal classification rule(CCR) is another set of rules which is applied when the closest node is labelled as border node and all the neighbours are from distinct classes. It allows the object to belong to specific classes along with their meta-classes with different masses of belief. The belief assignment of the object to a meta-class depends both on the distances to the centers of the specific class and on the distance to the meta-class center. Some objects too far from the others will be considered as outliers. CCR provides the robust classification results since it reduces the risk of misclassification errors by increasing the non-specificity. In the determination of the mass of belief on a meta-class U for X_s, the smaller distances between X_s and centers of the specific classes means that X_s is more likely to belong to this set of classes (i.e. U), and the partial imprecision degree of the class of X_s is higher when X_s is closer to the center of U (i.e. C_U)). So the mass of belief for X_s on the meta-class U should be a function denoted by $f_2(\hat{\mathrm{A}}\cdot)$ of both the distance to each center of the specific class involved in U and distance to center of meta-class C_U, which is given by:

$$m(U) = f_2(\Delta(X_s, C_U)) \tag{1}$$

According to these mass function these objects are classified under Credal classification.

B. Evolutionary Fuzzy ARTMAP Neural Networks for Classification of Semiconductor Defects

An industry need to identify the defective semiconductors during the processing step. But since the original dataset is very large and the defective items contribute only a little(minority class) there will be difficulties in classification. In addition, the database may comprise overlapping samples of different classes. A system was proposed[4] which introduces two models of evolutionary fuzzy ARTMAP (FAM)[5] neural networks to deal with the imbalanced data set problems in a semiconductor manufacturing operations. In particular, both the FAM models and hybrid genetic algorithms are integrated in the proposed evolutionary artificial neural networks (EANNs) to classify an imbalanced data set. The FAM structure is formed by a pair of fuzzy ART modules, interconnected by a map field. It places its input from one module to the target class in the other module through its layers normalization, input and recognition layer. It does a pattern matching process and identifies the defective modules. A hybrid genetic algorithm(HGA) is also integrated with FAM module. Figure 1 shows the architecture of FAM module. The idea of the two-phase HGA[6] is that, in the first phase it does a base level search algorithm for candidate solutions that are located in the basin of global optimum whereas the direct

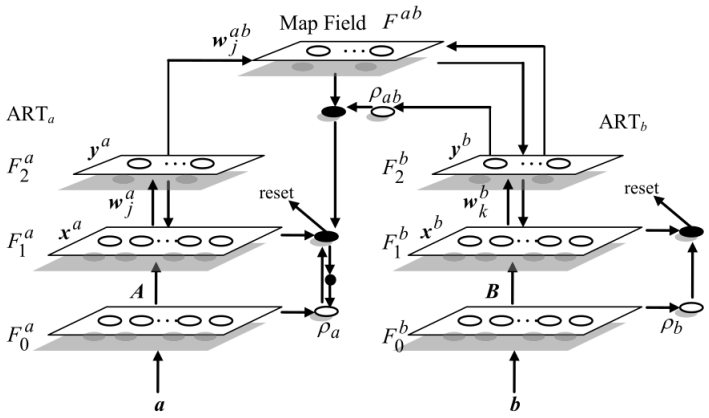

Figure 1. FAM module

search(DS) algorithm, which provides a systematic procedure to reduce the size of the search space for finding an exact optimal solution, is initiated in the second phase to conduct a local search by refining the best candidate solution found by the GA. Both FAM and HGA are combined for improving the efficiency of classification. The proposed FAM-HGA model is an evolutionary FAM ANNs that undergoes a learning process by a supervised ART in the network environment and an evolutionary search and adaptation process by HGA in an evolution environment. The HGA is deployed to enhance the learning capability of FAM by searching and adapting the network weights. In the network environment, the network operation is performed within the FAM.

C. A Hybrid Belief Rule-Based Classification System Based on Uncertain Training Data and Expert Knowledge

Another system was proposed [7] for pattern classification. This system is more applicable when complementary informations[8] are available. In some cases, we may have training data obtained from sensors and some expert knowledge obtained through some interviews by experts. Both of these datas are considered in this classification. Training data and knowledge are first coded into rules or relations and then inference is made accordingly. The production rules (especially, the fuzzy IF-THEN rules) are selected to represent expert knowledge. That is, experts are asked to assign fuzzy regions to each class and to give corresponding certainty grades. When each piece of expert knowledge is represented and then it is expande to belief rules. These rules are then combined according to their antecedent part, consequence part and rue weight. As in figure 2, both KBRB and DBRB are then fused by giving some weights for each type of data. Due to the independence between training data and expert knowledge, the fuzzy regions covered by the DBRB and the KBRB does not fully overlap. Thus, the rules in the HBRB can be divided into the three categories: rules with fuzzy regions only covered by the DBRB, rules with fuzzy regions only covered by the KBRB, and rules with fuzzy regions covered by both the DBRB and the KBRB A decision making strategy is built by calculating beliefs, plausibility and pignistic

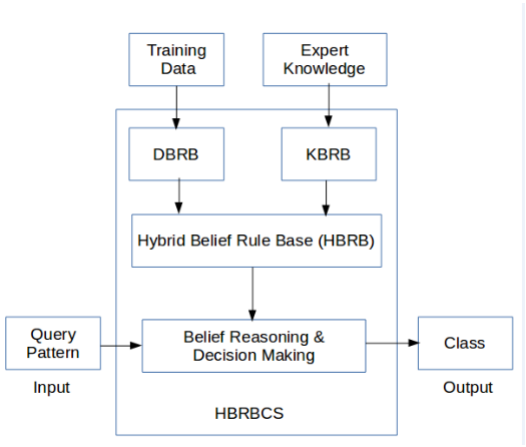

Figure 2. Architecture

probability. These are the useful standards of measurement. Mostly the user wishes a single class for the object using hard strategy. But in some cases soft strategies need to be adopted classify objects to providing multiple decision objects for further analysis. This will reduce errors. The belief function theory provides a better tools to develop both the hard and the soft decision strategies.

D. A Hybrid Static-Dynamic Classification for Dual-Consistency Cache Coherence

Cache consistency and coherence is an important aspects in operating system memory management. So managing these consistency is a difficult task since it first needs the identification of the regions which are data race free(DRF) and non data race free(nDRF). A technique was proposed[9] provides a dual-consistency cache coherence protocol that supports two execution modes: a traditional sequential-consistent protocol[10] and a protocol that provides weak consistency[11]. This does a static-dynamic hybrid classification of memory accesses based on (i) a compile-time identification of xDRF code regions for and (ii) a runtime classification of accesses based on the OSâs memory page management. First it perform a compile-time identification of extended data-race-free code region, where each xDRF region consists of a set of DRF regions. Then complement the static classification per regions with a dynamic private shared classification of memory accesses, by resorting to the OSâs memory management. The hybrid classification includes three protocol modes that are designed to handle each memory access: (i) OS private memory accesses require minimum coherence support and are optimized, (ii) accesses performed within xDRF regions can be executed under a high-performance and scalable Sequential Consistency for DRF protocol and finally (iii) for accesses that are neither OS-private nor part of xDRF regions, coherence is ensured by a standard directory protocol, which is commonly optimized for executing racy code. The static classification scheme focus on classifying memory accesses as private or shared based on the nature of the accessed data. The compile-time classification is complemented by a standard

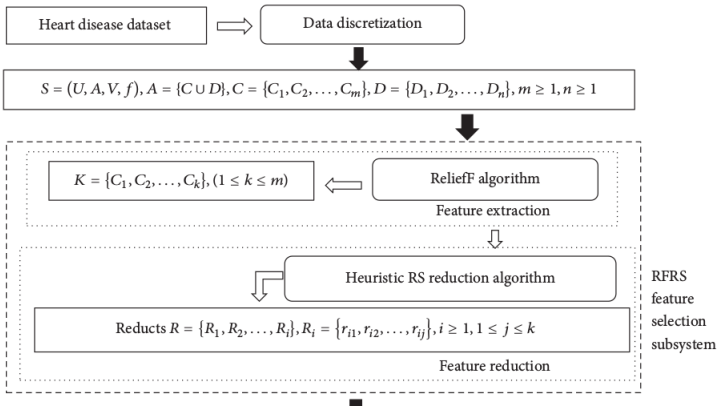

Figure 3. Feature selection subsystem

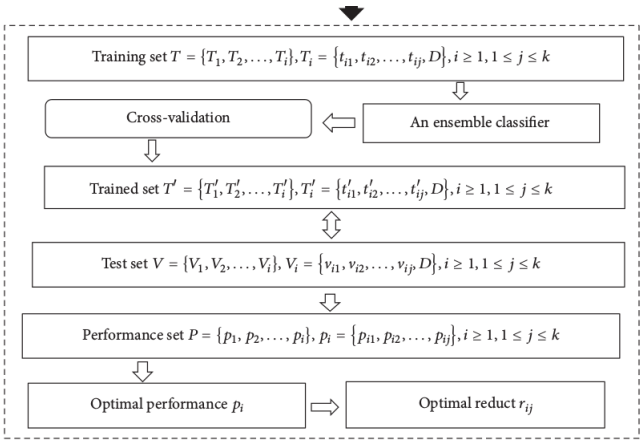

Figure 4. Classification subsystem

OS-based classification to increase the accuracy.

E. A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method

Another system proposed[12] was to aid the diagnosis of heart disease based on the ReliefF and Rough Set (RFRS) method. It comprises of two subsystems. A feature selection subsystem and a classification subsystem. Feature selection subsystem is based on RFRS method. It is hybrid feature selection system which employs reliefF[13] and roughset method for feature selection. reliefF algorithm is used to extract only the relevant features and Rough Set approach is used for heuristic reduction of the feature set.

After extracting the features, these are feeded to classification subsystem as shown in figure 3. Here the dataset is split into training sets and corresponding test sets. It uses an ensemble classifier boosting[14] algorithm with C4.5 as base classifier.

V. DISCUSSIONS AND FINDINGS

From the above works, I infer that some systems does not support dynamic data streams whereas some are restricted to some platforms. For a better classification, we have to solve

these problems also. Thus we can extend the work by adding some extra functionalities to support dynamic data streams. Also since feature selection is an important task in classification, a hybrid technology can be adopted for feature selection also. Thus we can ensure only relevant datas are chosen and these data can be classified to appropriate labels which will be the perfect class for that object.

We have seen that these systems do have some drawbacks. In the first system[1] it shows less computational complexity due to the of self organizing map. Though the object space is large enough to calculate the distance vectors of each object, it maps every object to a lower dimensional space, thus reducing the time complexity. But the negative side of using SOM is that it cannot auto determine the number of nodes in SOM.In the second system[4] it overcomes the stability plasticity dilemma thus enhancing the performance of classification. But since it requires the multiple vigilence test, there comes the computational complexity. As the number of input increases to match with the target pattern the number of times the vigilence test to be conducted also increases. This increases the overall computation complexity. Now in the third system[7] it gives a better performance since we are taking complementary informations. It gives the best result by considering both trained data and expert knowledge. Its main limitation is that it is not applicable to the classification of dynamic data streams. Next system[9] is widely used for the classification of DRF codes due to its minimum execution time. But the problem with this technique is that it is suited on OpenMP platform only. The final system[12] is very useful in diagonosing heart disease. There the efficiency is mainky due to hybrid feature selection methods by using reliefF and Roughset algorithms. Since the features are perfectly extracted and they reduced to minimum, computation complexity is also reduced. The main problem with this technique is that the number of nodes and weights[15] are not stable.

So here the five systems described above have their own limitations. When it shows its advantage in one direction, it has disadvantages in the other. Also these are all applicable in different environment. When one system works well in a particular case, the other may not. So the technique should be adopted carefully by looking at the needs and the situations. The table shown in I is a comparison of different works we have discussed above. These systems are used to classify different sets of data. That is mainly uncertain or imbalanced datasets. Also data can be training data that is obtained from sensors or expert knowledges. Also the data may be code regions or some areas. These are all different types of data which needs classification under certain circumstatnces.

TABLE I
COMPARISON ON DIFFERENT WORKS

Reference	Method Used	Type of Data
[1]	Hard Fuzzy and Credal classification	Uncertain data
[9]	HBRBCS with DBRB and KBRB	Complementary data
[4]	FAM and HGA	Imbalanced data set
[10]	Static and Dynamic	DRF codes and PS data
[14]	RFRS and Ensemble	Imbalanced dataset

VI. CONCLUSION

Classification problem is one of the most fundamental problem in data mining literature. Here we had text data, so the problem can be considered as the classification of set valued attributes. These new methods contribute high classification capabilities by combining the strengths of individual computing techniques. The shortcomings of individual techniques are compensated. There is nothing like best algorithm. Each algorithm plays as the best for one or the other cases. Here we have seen five different ways to combine different classification modules to enhance the performance of object classification. Five of these methods are applicable in different environment. All these methods have some limitations. Some does not support dynamic data streams whereas some are restricted to some platforms. Thus we can extend the work by adding some extra functionalities to support dynamic data streams. Also since feature selection is an important task in classification, a hybrid technology can be adopted for feature selection also. Thus we can ensure only relevant datas are chosen and these data can be classified to appropriate labels which will be the perfect class for that object.

REFERENCES

[1] Zhun-Ga Liu, Quan Pan, Jean Dezert, and Gregoire Mercier, "Hybrid Classi
cation System for Uncertain Data" IEEE Transactions on Systems, Man and Cybernetics: Systems, 2016.

[2] A. K. Agrawala, Ed.,1977 "Machine Recognition of Patterns". New York, NY, USA: IEEE Press.

[3] I. Hammami, G. Mercier, A. Hamouda, and J. Dezert, oct 2016, âKohonenâs map approach for the belief mass modeling,â IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 10, pp. 2060â2071.

[4] Shing Chiang Tan, Junzo Watada, Member, IEEE, Zuwairie Ibrahim, and Marzuki Khalid, "Evolutionary Fuzzy ARTMAP Neural Networks for Classification of Semiconductor Defects", IEEE Transactions on Neural Networks and Learning Systems, Vol.26, No. 5, May 2015.

[5] G. A. Carpenter, S. Grossberg, N. Markuzon, J. Reynolds, and D. Rosen, sep 1992, âFuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps,â IEEE Trans. Neural Netw., vol. 3, no. 5, pp. 698â713.

[6] S. Baskar, P. Subraraj, and M. V. C. Rao, dec 2001, âPerformance of hybrid real coded genetic algorithms,â Int. J. Comput. Eng. Sci., vol. 2, no. 4, pp. 583â602.

[7] Lianmeng Jiao, Thierry Denoeux, Member, IEEE, and Quan Pan, Member, IEEE, "A Hybrid Belief Rule-Based Clas-

sification System Based on Uncertain Training Data and Expert Knowledge",IEEE Tansactions on Systems, Man and Cybernetics: Systems, 2015.

[8] D. Dubois, P. HÃ¡jek, and H. Prade, 2000, âKnowledge-driven versus datadriven logics,â J. Logic Lang.Inf., vol. 9, no. 1, pp. 65â89.

[9] Alberto Ros and Alexandra Jimborean, "A Hybrid Static-Dynamic Classification for Dual-Consistency Cache Coherence", IEEE Transactions on Parallel and Distributed Systems, 2015.

[10] L. Lamport, âHow to make a multiprocessor computer that correctly executes multiprocess programs,â IEEE Transactions on Computers (TC), vol. 28, no. 9, pp. 690â691.

[11] D. J. Sorin, M. D. Hill, and D. A. Wood, 2017 "A Primer on Memory Consistency and Cache Coherence, ser. Synthesis Lectures on Computer Architecture", M. D. Hill, Ed. Morgan Claypool Publishers, 2011.

[12] Xiao Liu, Xiaoli Wang, Qiang Su, Mo Zhang, Yanhong Zhu, Qiugen Wang, and Qian Wang, "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method", Computational and Mathematical Methods in Medicine, Volume 2017.

[13] L.-X. Zhang, J.-X. Wang, Y.-N. Zhao, and Z.-H. Yang, âA novel hybrid feature selection algorithm: using ReliefF estimation for GA-Wrapper search,â in Proceedings of the International Conference onMachine Learning andCybernetics, vol. 1, pp. 380â384, IEEE, Xiâan, China, November 2003.

[14] J. Sun, M.-Y. Jia, and H. Li, 2011 âAdaBoost ensemble for financial distress prediction: an empirical comparison with data from Chinese listed companies,â Expert Systems with Applications, vol. 38, no. 8, pp. 9305â931

[15] D. Wettschereck, D. W. Aha, and T. Mohri, 1997, âA review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms,â Artificial Intelligence Review, vol. 11, no. 1â5, pp. 273â314.