

Comparative Study of Different Classification Techniques: Heart Disease Use Case

Hanane Bouali

Bestmod, Institut Supérieur de Gestion
Tunis, Tunisia
hanene.bouali@gmail.com

Jalel Akaichi

Bestmod, Institut Supérieur de Gestion
Tunis, Tunisia
j.akaichi@gmail.com

Abstract--Common stream mining tasks include classification, clustering and frequent pattern mining among them; data stream classification has drawn particular attention due to its vast real-time application. Through these applications, the main goal is to efficiently build classification models from data streams for accurate prediction. The development of such model has shown the need for machine learning techniques to be applied to large scale data. A range of machine learning techniques exists and the selection of the accurate techniques is based on advantages and limits of each one and how these latter well addresses important research techniques. In this paper, we present the comparison of different classification techniques using WEKA in order to investigate the performance of a collection of classification algorithms. This comparison shows the support vector machine performance with higher accuracy and better results when classifying our dataset.

Keywords-- *Dynamic System; Classification; Machine Learning techniques; Heart Disease; WEKA*

I. INTRODUCTION

A major problem in bioinformatics analysis or medical science is in attaining the correct diagnosis of certain important information. For the ultimate diagnosis, many tests generally involve the classification of large dataset. However, too many tests may complicate the main diagnosis process and lead to difficulty in obtaining and interpreting end results. This kind of difficulty may be solved using the machine learning.

Machine learning techniques can extract flexible and comprehensible knowledge from large dataset. They also require knowledge for their effective use, but are less complicated to employ and their results are more comprehensible to users. Machine learning techniques can deal with a mix of quantitative, qualitative, missing or noisy data so common on engineering.

Machine learning algorithms are described as either supervised or unsupervised. The distinction is drawn from how the learner classifies data. In supervised learning, which referred also to classification, the classes are predetermined. These classes can be conceived of as a finite set, previously arrived at by human. In practice, a certain segment of data will be labeled with these classifications. The machine learner's task is to search for patterns and construct models. These models then are evaluated on the basis of their

predictive capacity in relation to measures of variance in the data itself.

Scientists use classification system to help them make sense of the world around them. They use classification to organize information and objects. When things are stored into groups, it makes them easier to understand and it makes it easier to see the relationships between them.

Classification's algorithms have used several techniques among them we find artificial neural networks, support vector machine, Bayesian network, fuzzy pattern trees and decision trees.

In this paper, we will discuss in details the techniques used to cope with classification issues and provides a comparative study based on the research issues. Besides the fundamental review on the selected techniques, we present an experimental evaluation using WEKA.

The remainder of this article is organized as follows:

In section 2 we present the research issues that every algorithm must cope with. In section 3, we present a literature review of algorithms handling classification problem while highlighting their objectives, features, complexity, and limits. In section 4, we provide a comparative study between techniques mentioned before. In section 5, an experimental evaluation is made using WEKA applied to dataset. We conclude in section 5 and provide future works.

II. RESEARCH ISSUES

Data stream classification in such real world application is typically subject to some major challenges. In this section, we will discuss research limits have been found in the context of data stream classification.

- High speed nature of data streams: the algorithm should be able to adapt to the high speed nature of the information. Also, we have the constraint of the one-pass; this refers to scan the data only once.
- Unbounded memory: to build the classification model, data need to be resident in memory. Hence, the huge amounts of data streams generated daily need an unbounded memory.
- Concept drifting: Concepts drifts change the classifier results over time. It is also referred as data stream evolution. With this evolution, many temporal relations can be generated or new classes. Hence, classifications algorithms must capture such change. This is guaranteed by the use of an outdated mode. The detection of such changes is in order to increase the classification accuracy.

Figure 1 illustrates the concept drifting in data streams. In the three consecutive time stamps T_1 , T_2 and T_3 , the classification boundary gradually drifts from b_1 to b_2 and finally to b_3 .

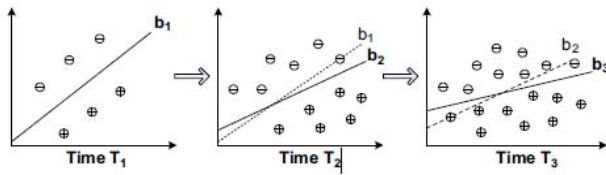


Figure 1. An illustration of concept of drifting in data streams [1]

- Hardware evolution: advances in hardware technology have allowed us to automatically record transactions and other pieces of information of everyday life at a rapid rate. Such process generates huge amounts of data streams.
- Partial labeling: due to large volume of stream data, it is unfeasible to label all stream for building classification model. Thus, data stream contains both labeled and unlabeled data.
- Real time accuracy evolution: in some cases, the user is not interested in mining data stream results, but how these results are changed over time. These changes can be in the number of initial classes, it might represent some changes in the dynamics of the arriving stream leading to changes in the knowledge structure. This especially in applications based on temporal analysis.
- The use of produced rules
- Disparate sources

III. CLASSIFICATION TECHNIQUES

There are several classification techniques for classification of both multivariate and univariate dataset. Some more recent approaches to classification are Artificial Neural Network (ANN), Support Vector Machine (SVM), Bayesian network (BN), Fuzzy Pattern Trees (FPT) and Decision Trees (DT). In this section, we present the before-mentioned techniques and provide some inconvenient and advantages of those techniques.

A. Artificial neural Network

ANN has been used in many fields with positive results ranging from pattern recognition to diagnostic classification task from character recognition to speech identification and from images processing to robotics. Managers and planners have long sought decision making tools for detecting changes in data streams over large spatial and temporal scales. The ANN system architecture is sufficiently flexible to allow for its continual update and refinement of new data. Their use has potential to save time and resources between input predictor and known output responses. Their ability to adapt continuously to new data allows them to track changes in a signal over time and their adaptability to learn from arbitrary, noise data permits to solve problems that cannot be handled adequately with some of the conventional statistical technique [2].

In addition to those characteristics, ANN are non-parametric classifiers so they are capable of classifying multi source data. Authors in this research [3] take advantages from these characteristics and apply this

technique to assimilate large amounts of disparate data types for use in fluvial hazard management decision making. Two types of ANN are used (kohenen self-organizing map and a counter propagation algorithm) in hierarchy to predict reach scale stream geomorphic condition, inherent vulnerability and sensitivity to adjustments. The use of ANN allows for an adaptive watershed management approach, does not require the development of site specific, physical based, stream models and provides a standardized approach for classifying river network sensitivity in various context.

This paper present modeling instable stream; such instability is due to varying types of watershed stressor magnitude, duration and periodicity in surface water quality. Hence, ecosystem managers are faced with the challenge of integrating data from disparate sources regarding their features. ANN used in this work to predict channel conditions (instability), offers many advantages including the ability to accommodate both large amounts of spatial and temporal data as well as multiple data types. It also provides a standardized, expert trained approach for classifying the sensitivity of river networks in various contexts.

Some characteristics of the neural network approach have been tested and validated for the particular problem of diagnostic classification in the field of computerized electrocardiography (Possibilities of using neural networks for ECG classification). Two different databases have been used for the evaluation process: CORDA, developed by the Medical Informatics Department of the University of Leuven, and ECG-UCL, developed by the Cliniques Universitaires Saint-Luc, Université Catholique de Louvain. Electrocardiographic signals classified on the basis of electrocardiographic independent clinical data, with a single diagnosis and no conduction abnormalities have been considered. Seven diagnostic classes have been taken into account, including the different locations of ventricular hypertrophy and myocardial infarction. Two architectures of neural networks have been analyzed in detail considering three aspects: the normalization process, pruning techniques, and fuzzy preprocessing by the use of radial basis functions.

B. Support Vector Machine

When using SVM, first transforming data into high dimensional space may convert complex classification problem into simpler problem that can use linear discriminant function. Secondly, SVM provides the most useful information for classification. When applying the SVM to a linear classification algorithm, we have to construct the SVM where the decision surface used to classify a pattern as belonging to one of the two classes in the hyper plan. But, in real conditions, data observed are frequently affected by outliers, caused by noisy measurements. If outliers are taken into account, the margin of separation decreases and hence, the solution does not generalize as well.

To cope with this problem, authors introduce costs functions to labeled points in order to have asymmetric soft margin [4]. On other side, SVM is faced to the problem of binary classification unlike problems in real world. Seeing when the number of classes is a power of two, authors in some researches adopts the binary tree with a SVM at each node [5]. The performance of the SVM is mostly attributed to the user's ability to introduce knowledge about data

unbalance and class confusion. To cope also with the problem of binary classifications, a DAG multi classification schema was chosen to extend SVM to the multi classification problem. A 5-fold cross validation was applied to find the optimal SVM hyper parameters.

Hidden Markov Model (HMM) topologies are also associated with SVM to explore the possibility of pre-segmenting the data [6] with a simple HMM before applying the SVM. These topologies are introduced using k-variable k-means algorithm.

Comparing the two approaches, we observe that the system based on SVM obtain better results than the system based on the HMM technology.

C. Bayesian Network

Bayesian network (BN) are parametric classifiers and have problems with multi source data. Sungbo et al. have proposed a new classification models that exploit temporal relations among features which within and across data stream. [7]. The consideration of such new relations improve significantly the classification accuracy. It also improves the interpretability of the resulting models. Data streams are collected using diverse sensors. These multiple diverse sensors are used to monitor changes. Sensor readings are obviously dependent. Hence, changes detected in one sensor might affect readings in others. However, the interdependency among sensor readings and their temporal relations have not been treated impervious classifications work.

To handle the problem of temporal relations, authors provide a monitoring scenario of mobile robots with many sensors. The robot is engaged in various tasks and sends to a central node.

D. Fuzzy Pattern trees

Fuzzy Pattern Trees (FPT) have been introduced as a novel approach class for machine learning [8]. Authors in this paper consider the problem of learning fuzzy pattern trees for binary classification from data streams. This approach takes into account new data items as soon as they arrive and develop a learning algorithm adaptive in the sense that an up to date model is offered at any time. This is ensured by anticipating possible local changes of the current model and confirming these changes through statistical hypothesis testing. A FPT is a hierarchical, tree like structure, whose inner nodes are marked with generalized logical and arithmetic operators, whereas the leaf nodes are associated with fuzzy predicates on input attributes.

Most existing works on classification of data streams assumes that all streaming data are labeled and the class labels are immediately available. However, in real world applications, this assumption is not always valid [9]. With this motivation, authors propose a semi-supervised classification algorithm for data stream with concept drifts and unlabeled data. To cope with those issues, the algorithm adopts the SVM techniques.

First, they generate an incremental decision tree using the incoming streaming data. Meanwhile, they develop the k-modes clustering algorithm to label unlabeled data.

E. Decision Trees

A decision tree (DT) is a supervised classifier that recursively partitions a data set into smaller subdivisions based on a set of simple tests at each internal node in the tree

[11]. The leaf nodes represent the class labels y_i . Training data set is used to learn the split conditions at each internal node and to construct a decision tree. For each new sample (i.e., feature vector x), the classification algorithm will search for the region along a path of nodes of the tree to which the feature vector x will be assigned. That is, the classification of a region is determined by a path from the root node to a leaf node.



Many extension of the DT exists. To learn a classifier, the CVFDT algorithm requires input of precise and fully labeled samples, which is impractical in many real world applications. The streaming data often contains uncertainty due to various reasons, such as imprecise measurement, missing values and privacy protection; the problem of classifying uncertain data streams with only positive and unlabeled samples has not been studied by the research community yet.

To cope with this problem, authors in [10] transform CVFDT and propose a novel algorithm namely puCVFDT. Experiments in real life world applications show that the proposed algorithm has strong capabilities to learn from uncertain data streams with positive and unlabeled samples and tackle concept drifts.

IV. COMPARATIVE STUDY

A variety of methods such as decisions trees (CART, C4.5...), artificial neural network, Bayesian network and variant of the kernel methods (adaptive kernels...) are used to cope with the classification problem.

A comparative study between techniques presented previously is done based on the research issues mentioned previously. This comparative study is summarized in the table below:

Techniques used 	BN	SVM	ANN	FPT	DT
Issues Handled 					
Concept drifts	*	*		*	*
High speed nature of data streams	*	*	*		
Real time accuracy changes				*	*
Disparate sources			*		
Hardware evolution				*	
Unbounded Memory			*		

V. EXPERIMENTAL EVALUATION

To gauge and investigate the performance on the selected classification methods or algorithms namely ANN, SVM, RB, DT, and FPT. The 75% data is used for training and the remaining is for testing purposes.

A. Methods

In WEKA, all data is considered as instances and features in the data are known as attributes. The simulation results are partitioned into several sub items for easier analysis and evaluation. On the first part, correctly and incorrectly classified instances will be partitioned into numeric and percentage value subsequently Kappa static,

mean absolute error and root mean squared error will be in numeric value only. We also show the relative absolute error and root relative squared error in percentage for references and evaluation.

B. Data Set Information

The classification task in this database is to determine the presence of heart disease in the patient [12]. It is integer valued from 0 (no presence) to 4. The database contains 14 attributes. All attributes are numeric valued

Attribute information:

1. Age: age in years
2. Sex: 1= Male; 0=Female
3. cp: chest pain type
 - Value 1: typical angina
 - Value 2: atypical angina
 - Value 3: non-anginal pain
 - Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
 - Value 0: normal
 - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
 - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
 - Value 1: upsloping
 - Value 2: flat
 - Value 3: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num: diagnosis of heart disease (angiographic disease status)
 - Value 0: < 50% diameter narrowing
 - Value 1: > 50% diameter narrowing(in any major vessel: attributes 59 through 68 are vessels)

C. Results and Discussion

The implementation in WEKA will be done by taking 5 different algorithms mentioned before. To analyze and compare those algorithms, the following parameters are used:

- Kappa statistic (KS) [13]: Kappa statistic is a generic term for several similar measures of agreement used with categorical data. Typically it is used in assessing the degree to which two or more raters, examining the same data, agree when it comes to assigning the data to categories. Kappa is a measure standardized to lie on -1 and 1 scale where

complete agreement corresponds to $K = 1$, and lack of agreement corresponds to $K = 0$. A negative value of kappa would mean negative agreement.

- Mean absolute error (MAE) [14]: MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average i.e how close a predicted model to actual model.
- Root mean squared error (RMSE) [14]: RMSE is the difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors.
- Relative absolute error (RAE) [15]: is relative to a simple predictor, which is just the average of the actual values. In this case, though, the error is just the total absolute error instead of the total squared error. Thus, the relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor.
- Root relative squared error (RRSE) [15]: is relative to what it would have been if a simple predictor had been used. More specifically, this simple predictor is just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the simple predictor. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted.

First we classify the dataset using two test options: using training method and 10 cross fold method. The test values of those parameters are given in table 1.

TABLE 1.

Classifier	Using training set method				
	KS	MAE	RMSE	RAE	RRSE
BN	0.5632	0.1914	0.3352	0.5184	0.7802
SVM	0.801	0.1541	0.1976	0.2317	0.5586
ANN	0.4467	0.2711	0.3682	0.7343	0.8571
FPT	0.797	0.1091	0.2518	0.2956	0.5860
DT	0.1971	0.3199	0.3999	0.8664	0.9309
Classifier	Use 10 cross fold method				
	KS	MAE	RMSE	RAE	RRSE
BN	0.4814	0.2188	0.3617	0.5925	0.8418
SVM	0.7651	0.1864	0.2897	0.4981	0.7612
ANN	0.3753	0.281	0.3919	0.7611	0.9121
FPT	0.4917	0.1946	0.3858	0.5270	0.8980
DT	0.1777	0.3295	0.4105	0.8923	0.9554

From the table 1 statistics, it's clear that the use of training set methods has better performance than the use of 10 cross fold method. To better see the results we present them graphical.

Figure 2.

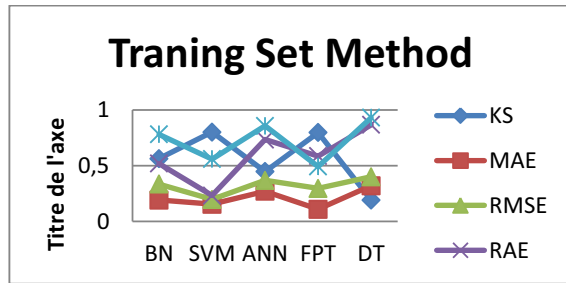


Figure 3.

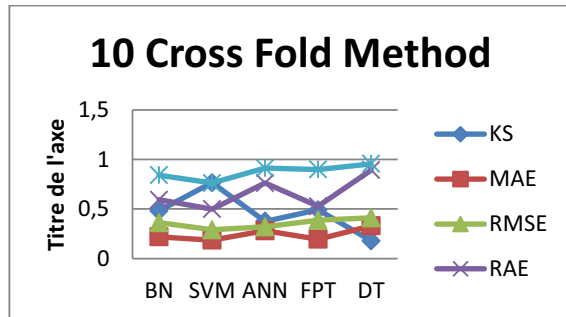


Figure 2 and 3 show that the comparison of classifiers with the help of two test options. These figures also states that use training set method has better performance than 10 cross fold method under these observations:

- The value of KS comparing above discussed method, use training set method has better value than 10 cross fold validation.
- Values of MAE and RMSE lower these values better the prediction. So use training set method has minimum values of AME and RMSE as compare to 10 cross fold validation.

From above analysis, we can now only consider use training set method to compare four algorithms.

TABLE II.

Algorithm	CCI (%)	ICI (%)	RAE (%)	RRSE (%)	Ts
BN	67.8284	32.1716	51.8425	78.0258	0.07
SVM	85.7655	14.2345	23.1765	55.8612	0.05
ANN	58.445	41.555	73.4316	85.7106	1.15
FPT	84.9866	15.0134	29.5617	58.6043	1.45
DT	42.8954	57.1046	89.2379	95.5408	0.01

Correctly classified instances(CCI)

Incorrectly classified instances(ICI)

Time taken in seconds

From the table above we can see the following:

- The correctly classified instances of Support vector Machine (SVM) and Fuzzy Pattern Trees (FPT) have higher values also for incorrectly classified instances.
- Time taken to build SVM model is better than FPT.
- The RAE and RRSE have almost same values for SVM and FPT but SVM has lower values.
- Decision tree has the worst values of KS and error but the best time taken to build the model.

From the above experiments, it's clearly seen that SVM and FPT are two best algorithms for classifying the patient dataset. But comparing these two algorithms the value of

kappa statistics SVM has higher values. So, it's clearly told that SVM is better than FPT for classifying our dataset.

VI. CONCLUSION AND FUTURE WORK

As a conclusion, we have met our objective which is to evaluate and investigate 5 selected classification algorithms based on Weka. The best algorithm based on the dataset is SVM.

These results suggest that among the machine learning techniques tested, Support Vector Machine has the potential to significantly improve the conventional methods for use in medical or in general, bioinformatics field.

REFERENCES

- [1] C.A Charu, "Data Streams: Models and Algorithms" (2nd Ed.). Springer, USA, 2007
- [2] B. Ajith, "Artificial Neural Network. John Wiley", USA 2005
- [3] L. Besaw,, D. Rizzo, , M. Kline, , K. Underwood, J. Doris, , L. Morrisseyand K. Pelletier, "Stream Classification using Hierarchical Artificial Neural Networks: A Fluvial Hazard Management Tool". Journal of hydrology, 2009, vol.373, pp.34-43
- [4] A. Temko, and C. Nadeu, "Classification of Acoustic events using SVM based Clustering Schemas Pattern recognition". 2009, Vol.39,pp.682-694
- [5] G. Guo and Z. Li, "Content-based audio Classification and Retrieval using Support Vector Machine". IEEE Trans. Neural Networks, 2003, vol.14, pp.209-215
- [6] G.M. Reyes and D. Elis, "Seelection, Parameter Estimation, and Discriminative Training of Hidden Markov models for General Audio Modeling". International Conference in Multimedia and Expo 2003
- [7] S. Sungbo, J. Kang, and H. Ryu, "Multivariable stream data classification using motifs and their temporal relations". Information System,2009, vol.179, pp.3489-3504
- [8] A. Shaker, R. Senge and E. Hüllermeier, "Evolving Fuzzy Pattern Trees for Binary Classification on Data Streams". Journal of Information Sciences,2013, vol.220, pp.34-45
- [9] X. Wu, P. Li, and X. Hu, "Learning from Concept Drifting Data Streams with Unlabeled Data". Neurocomputing,2012, vol.92, pp.145-155
- [10] C. Liang, Y. Zhang, P. Shi, and Z. Hu, " Learning Very Fast Decision Trees from Uncertain Data Streams with Positive and Unlabeled Samples". Information Sciences, 2012, vol.213, pp.50-67
- [11] Arff Datasets, <http://repository.seasr.org/Datasets/UCI/arff/breast-cancer.arff>
- [12] Glossary of statistical terms, <http://www.statistics.com/>
- [13] Mean Absolute error and root mean absolute error, <http://www.eumetcal.org/>
- [14] Analyzing GeneXproTools Models Statistically, <http://www.gepsoft.com/>