Techniques of Data Stream Mining for Health Care Application

[1]Satish Thombre , [2]Snehalata Dongre

[1]Department of Computer Science and Engineering, G.H.Raisoni College of Engineering, Nagpur, Maharastra, India

[2]Department of Computer Science and Engineering, G.H.Raisoni College of Engineering, Nagpur, Maharastra, India

Abstract

Data stream mining plays a key role to analyze the continuous data. The effective and efficient analysis of this data in such different forms becomes a challenging task. Developments in sensors device, miniaturization of low power micro-electronics device, and wireless networks devices are becoming a significant opportunity for good quality of health care services. Signals like ECG, EEG, and BP etc. can be monitor through wireless sensor networks and analyzed with the help of data mining techniques. These real-time signals are continuous in nature and abruptly changing hence there is a need to apply an efficient and concept real-time data stream mining techniques for taking intelligent health care decisions. The high speed and large amount of data set in data stream, the traditional classifier and classification technologies are no more applicable. The important criteria are to handle the 'concept drift' in data streams mining. Data Stream Mining (DSM) process performs data analysis and may uncover important data patterns, knowledge base data, and scientific and medical research data.

Keywords: *Concept drifts, Data Stream Mining.*

1. Introduction

Computer based methods for analysis of medical signals has been subject of intense research. Applications that perform signal processing and diagnostic interpretation of signals are widely spread. One example is a visual analysis of long term electrocardiogram (ECG) signal (sometimes called Holter signal) or analysis of sleep electroencephalogram (EEG) that is tedious time-consuming and operator dependent. ECG signals are biomedical signals originating from the actions of the human heart. The human heart operates cyclically pumping blood into the arterial system, which causes bioelectrical variations detectable from human skin. The action of the heart causing the ECG signal is depicted in Figure 1., where we have one real cardiac cycle of an ECG signal. The whole ECG recording consists of several consecutive cardiac cycles. In the cardiac cycle there are fairly periodic waves and peaks corresponding to the consecutive phase of the heart's action.

A cardiac cycle of an ECG signal includes a QRS complex as its primary and most dominant subpattern. Before the QRS complex there is a P wave in the ECG, and after the QRS complex a T wave, which is larger than the P wave. Flat segments between the above mentioned components are PQ, ST and TP segments. An RR interval, which is the delay between two consecutive QRS complexes, gives us useful information about the action of the heart. The most important parameters for the cardiologists are the durations and the amplitudes of the above mentioned subpatterns.

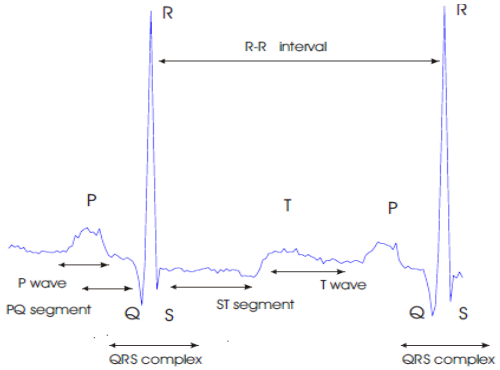

Figure 1.: Representative cycle of ECG signal.

2. Generation of the ECG signal

In case of an ECG signal analysis, the input string is either the amplitude or a broken line approximation of an ECG signal achieved by the method [6]. The first feature is the duration of the line segment t_i, and the second feature is the amplitude of the start point of the line segment h_i. The ECG signal 100 2 (the second lead) from MIT/BIT arrhythmia database whose length after segmentation was approximately 300 segments, T = 300. Therefore the input string can be described as follows.

$$O_i = \begin{bmatrix} t_i \\ h_i \end{bmatrix} \in \Re^2, 1 \leq_i \leq T \qquad 1.$$

When producing an artificial signal, we first choose one initial state according to the initial probability function.

3. Mining data stream techniques

A classifier is a mapping between a feature space and a label space, where the features represent characteristics of the elements to classify and the labels represent the

classes. For example, a Medical recommended system can be implemented by a classifier that classifies ECG signal into different categories based on a number of features that describe it. There are many types of classifiers, but in general we will talk about either supervised or unsupervised classification. In supervised classification, a set of labels or categories is known in advance and we have a set of labelled examples which constitute a training set. In unsupervised classification, the labels or categories are unknown in advance and the task is to suitably organize the elements at hand. In this section we describe several algorithms to learn supervised classifiers.

3.1 Nearest Neighbours

The nearest neighbour classifiers work by storing training records and using them to predict the class label of medical signal. This classifier memorizes the entire training set and classifies only if the attributes of the new record match one of the training examples exactly. A more elaborate, and far more popular, instance based classifier is the nearest neighbor classifier (kNN) [1]. given a point to be classified, the kNN classifier finds the k closest points (nearest neighbors) from the training medical records. It then assigns the class label according to the class labels of its nearest neighbours. The underlying idea is that if a medical record falls in a particular neighbourhood where a class label is predominant it is because the medical record (like ECG Signal) is likely to belong to that very same class.

Given a query point "q" for which we want to know its class "c", and a training set X = {{x1, c1}...{xn}}, where xi is the i-th element and c i is its class label, the k-nearest neighbors will find a subset Y = {{y1,c1}...{yk}} such that Y ⊆ X and

Σk 1 d(q,yk) is minimal. Y contains the k points in X which are closest to the query point q. Then, the class label of q is c = f ({c1...ck}).

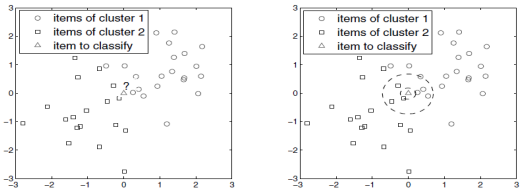

Fig. 2.: Example of k-Nearest Neighbours.

The left subfigure shows the training points with two class labels (circles and squares) and the query point (as a triangle).The right sub-figure illustrates closest neighbourhood for k = 1 and k = 7. The query point would be classified as square for k = 1, and as a circle for k = 5 according to the simple majority vote rule. Note that the query points were just on the boundary between the two clusters. Perhaps the most challenging issue in kNN is how to choose the value of k. If k is too small, the classifier will be sensitive to noise points. But if k is too large, the neighbourhood might include too many points from other classes. The right plot in Fig.2 shows how different k yields different class label for the query point, if k = 1 the class label would be circle whereas k = 7 classifies it as square. One of the advantages of this classifier is that it is conceptually very much related to the idea of CF (collaborative filtering): Finding likeminded users (or similar items) is essentially equivalent to finding neighbours for a given user or an item. The kNN approach, although simple and intuitive, has shown good accuracy results. The kNN method is difficult for medical data classification, because there is difficult to choose the value of k, if k is too small, the classifier will be to sensitive, and if the value of k is too large the neighbour might be include to many value.

3.2 Decision Trees

Decision trees [4,5] are classifiers on a target attribute (or class) in the form of a tree structure. The observations (or items) to classify are composed of attributes and their target value. The nodes of the tree can be: a) decision nodes, in these nodes a single attribute-value is tested to determine to which branch of the sub tree applies. Or b) leaf nodes which indicate the value of the target attribute. Once the partition has been found, the algorithm is recursively repeated until a partition is empty or all the observations have the same target value. Splits can be decided by maximizing the information gain, defined as follows,

$$\Delta_i = I(\text{parent}) - \sum_{j=1}^{ki} \frac{N(V_j) I(V_j)}{N} \qquad 2.$$

Where ki are values of the attribute i, N is the number of observations, v j is the jth partition of the observations according to the values of attribute i. Finally, I is a function that measures node impurity. The difficulty to use the decision tree is that, it only can analyze static and finite data set. Cannot handle data stream

3.3 Ruled-based Classifiers

Rule-based classifiers classify medical data by using a collection of "if . . . then . . ." rules. The rule antecedent or condition is an expression made of attribute conjunctions.

The rule consequent is a positive or negative classification. In order to build a rule-based classifier for medical data we can follow a direct method to extract rules directly from data. The advantages of rule-based classifiers are that they are extremely expressive since they are symbolic. But it is difficult for experts to transfer their knowledge into distinct rules, and it needs many rules to make system effectively.

3.4 Bayesian Classifiers

A Bayesian classifier [1,3] is a probabilistic framework for solving classification problems of medical data. It is based on the definition of conditional probability and the

Bayes theorem. The Bayesian method of statistics uses probability to represent uncertainty about the relationships learned from the data. In addition, the concept of priors is very important as they represent our expectations or prior knowledge about what the true relationship might be. Bayesian classifiers consider each attribute and class label as (continuous or discrete) random variables. Given a record with N attributes (A1, A2, ...,AN), the goal is to predict class Ck by finding the value of Ck that maximizes the posterior probability of the class given the data P(Ck|A1,A2, ...,AN). Applying Bayes' theorem, P(Ck|A1,A2,.,AN) = P(A1,A2,.,AN|Ck) P(Ck) A particular but very common Bayesian classifier is the Naive Bayes Classifier. In order to estimate the conditional probability, P(A1,A2, ...,AN|Ck), a Naive Bayes Classifier assumes the probabilistic independence of the attributes i.e. the presence or absence of a particular attribute with each other. P(A1,A2,...,AN|Ck)=P(A1|Ck)P(A2|Ck)...P(AN|Ck).

The main advantages of Naive Bayes classifiers are that, they are robust to isolated noise points and irrelevant attributes, and they handle missing values by ignoring the instance during probability estimate calculations. However, the independence assumption may not hold for some attributes as they might be correlated. Difficulty to get the probability knowledge for possible diagnosis and not being practical for large complex systems given multiple symptoms

3.5 Artificial Neural Networks

An Artificial Neural Network (ANN) [2] is an assembly of inter-connected nodes and weighted links that is inspired in the architecture of the biological brain. Nodes in an ANN are called neurons as an analogy with biological neurons. These simple functional units are composed into networks that have the ability to learn a classification problem after they are trained with sufficient data.

The simplest case of an ANN is the perceptron model, illustrated in figure 3. If we particularize the activation function to be the simple Threshold Function, the output is obtained by summing up each of its input value according to the weights of its links and comparing its output against some threshold θk. The output function can be expressed using Eq. 3.

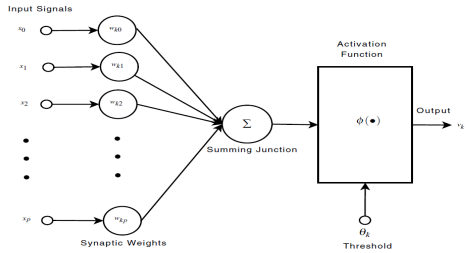

Fig. 3.: Perceptron model

The perceptron model is a linear classifier that has a simple and efficient learning algorithm. But, besides the simple Threshold Function used in the Perceptron model, there are several other common choices for the activation function such as sigmoid, tanh, or step functions.

$$Y(k) = \begin{cases} 1 \text{ if} \sum xiWki \geq \theta k \\ 0 \text{ if} \sum xiWki \leq \theta k \end{cases} \qquad 3.$$

The main advantages of ANN are that depending on the activation function they can perform non-linear classification tasks, and that, due to their parallel nature, they can be efficient and even operate if part of the network fails. The main disadvantage is that it is hard to come up with the ideal network topology for a given problem and once the topology is decided this will act as a lower bound for the classification error. The difficulty to use this method to classify the medical data is that the training process consume so much time that users cannot use the systems effectively

4. Conclusion

We studied the data mining technique as well as we discuss a comprehensive classification technique for data stream mining based on data mining applications. Here introduce the different type of data stream classification technique. These techniques are not much more helpful to classify large amount of medical data, so we use new technique to classify the large amount of data and give the correct accuracy.

References

[1] Remco R. Bouckaert. Voting massive collections of bayesian network classifiers for data streams. In Abdul Sattar and Byeong Ho Kang, editors, Australian Conference on Artificial Intelligence, volume 4304 of Lecture Notes in Computer Science, pages 243–252. Springer,2006.

[2] Zurada, J., Introduction to artificial neural systems. West Publishing Co., St. Paul, MN, USA, 1992.

[3] Friedman, N., Geiger, D., and Goldszmidt, M., Bayesian network classifiers. Mach. Learn., 29(2-3):131–163, 1997.

[4] Quinlan, J.R., Induction of decision trees. Machine Learning, 1(1):81–106, March 1986.

[5] Rokach, L., Maimon, O., Data Mining with Decision Trees: Theory and Applications, World Scientific Publishing (2008).

[6] Daniel Nov´ak, Electrocardiogram Signal Processing using Hidden Markov Models, 2003

[7] Leo Breiman. Pasting small votes for classification in large databases and on-line. Machine Learning, 36(1-2):85–103, 1999.

[8] J. Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[9] J. Ross Quinlan. Induction of decision trees.Machine Learning,1(1):81–106,1986.

[10] Ludmila I. Kuncheva. Classifier ensembles for changing environments.In Fabio Roli, Josef Kittler, and Terry

Windeatt, editors, Multiple Classifier Systems, volume 3077 of Lecture Notes in Computer Science, pages 1–15. Springer, 2004.