Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

7th International Conference
on Electrical Engineering
ICEENG 2010

Pattern-based Data-Classification Technique

By

M. Salama* A. Hasanen** A. Fahmy***

Abstract:

This paper presents a novel model of a supervised machine learning approach for classification of a dataset. The model depends on a feature selection (dimensionality reduction) method that is based on pattern-based subspace clustering technique. Then this clustering technique is applied to the dataset to perform the classification of the data. This approach is a non-statistical technique that supports most of the requirements that have been discussed recently like dimensionality reduction using multivariate feature selection method, threshold independence and handling of missing data. The approach tends to handle these requirements altogether which not the case in other classification models as discussed in this paper. Another distinguishing point in this model is its dependence on the variation of the values of relative features among different objects. Experimental results on synthetic and real datasets show that approach outperforms the existing methods in both efficiency and effectiveness.

Keywords:

Feature Selection, Classification, Patterns.

---

* Egyptian Armed Forces
** Benha High Technology Institute, Benha, Egypt
*** College of Engineering, Cairo University, Cairo, Egypt

1. Introduction:

Data mining tasks can be classified into two categories: descriptive like clustering techniques and predictive like classification techniques. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions [1]. Classification is a widely used technique in various fields, including data mining, whose goal is to classify a large set of objects into predefined classes, described by a set of attributes, using supervised learning methods. Classification algorithms output formats are a knowledge representation schemes like decision rules and decision trees. These decision rules are then applied to classify the objects [2]. Examples of basic techniques for data classification like how to build decision tree classifiers, Bayesian classifiers, Bayesian belief network, rule based classifiers, Back propagation (a neural network technique) and more recent approaches to classification like support vector machines, genetic algorithm and fuzzy logic techniques [3, 4]. Many attempts have been made in the last decades to design hybrid systems for pattern classification by combining the merits of individual techniques [5]. Most approaches perform dimension reduction as a preprocessing of the data then apply classification method afterwards. Feature selection is one of the important and frequently used techniques in data preprocessing of data mining algorithms [6]. Feature Selection is a process of an attribute selection method in decision tree for splitting attributes (find the features the best divide the training data) then used in some studies to reduce the high dimensionality of the feature space [8]. More about Dimensionality reduction will be discussed later in this paper.
On the other hand, clustering analysis is a procedure to partition a set of objects into a number of subsets (each subset is referred to as a cluster), each of which contains only objects as similar as possible, based on a certain pre-specified similarity metric. Up to date, a number of methods have been developed to find clusters in full feature space [9]. These approaches are effective and efficient over low-dimensional datasets. However, as the dimension increases, their performance deteriorates sharply due to the curse of dimensionality, to overcome this difficulty, we may consider using feature (or attribute) selection techniques like frequent pattern-based clustering [3]. A number of recent approaches adopted a semi-supervised model for classification. These approaches first apply unsupervised, flat clustering algorithms (k-mean clustering) to cluster all instances (training and testing data) in the dataset, and then use the resulting clustering solution to add additional instances to the training set [10]. Also this movement from clustering to classification model appear in Learning Vector Quantization (LVQ) which is based on a standard Self Organizing Maps with input vectors {x} and weights vector (representatives) {wj} where input data points have associated class information. This

allows us to use the known classification labels of the inputs to find the best classification label for each wj. SOM algorithm aim to map continuous input space into a low dimensional spatially discrete output space (feature map). SOM algorithm first set an initial weight vectors wj and for each input sample x of D features, the Euclidean distance dj(x) is calculated as in formula (1)

$$d_j(x)=\sum_{i=1}^{D}(x_i - w_{ji})^2 \quad (1)$$

then for weight wj with the minimum dj(x), the corresponding weight is updated by the following formula (2)

$$w_j= w_j + \eta(t)*(x_j - w_{ji})^2 \quad (2)$$

This procedure repeated iteratively (with a decreasing learning rate η(t)) until the map is become not changing [11]. The standard LVQ has drawbacks like the instability behavior in the case of overlapped data and the strong dependence on the initial positions of the representatives [12].

The idea of LVQ approach in transition from clustering to classification model is similar to the approach discussed in this paper. The approach is a classification model that is based on a standard clustering technique which pattern-based clustering technique. The new factor is that the input data have associated class information. This allows us to use the known classification labels of the inputs to find the best classification label for each pattern. There are many techniques that is based in clustering but they failed to support some of the requirements of classification as discussed in this paper.

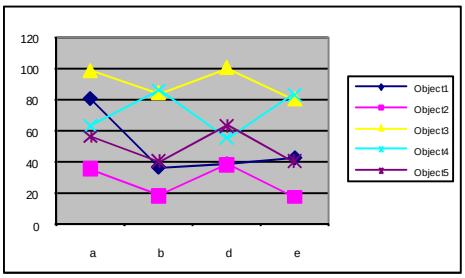
In the next section pattern-based subspace clustering is explained. In section 3 and 4, Dimensionality reduction methods and missing data techniques are discussed to show which type we are going to use in our approach. The problems we need to handle in the proposed approach is declared in section 5. The section 6 and 7 shows the proposed model and the experimental results. A discussion about the contribution of proposed model and the future work is explored in section 8

2. Pattern-based Subspace Clustering:

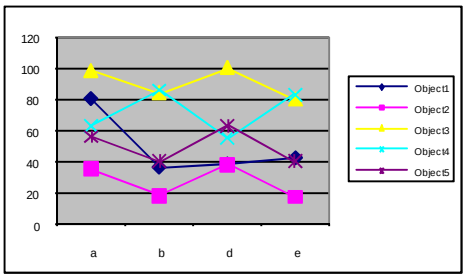
To illustrate pattern-based clustering, we give an example in Fig. 1. Fig. 1a is a dataset consists of five objects with five attributes. Fig. 1b shows the values of the objects in full space (five attributes), where no obvious pattern is visible. However, if we just select attributes {a, b, d, e} as in Fig. 1c for objects {2, 3, 5}, we can observe the following pattern: for all the three objects, from attribute a to attributes b; d and e, the values first go down, and then up and finally down. We can assign these three objects into the same subspace cluster as they show similar pattern. Likewise, similar patterns may exist with other objects in other subspaces [3].

Attribute / Objects	a	B	C	D	e
1	80	36	55	38	42
2	35	18	26	38	17
3	98	84	45	100	80
4	63	86	72	55	83
5	56	40	50	63	40

(a) The dataset



(b) Data in Full Space



(c) Pattern in subspace

Figure (1): *An Example of pattern-based clustering*

To tell whether two objects in D exhibit a coherent pattern in a given subspace S, it is essential to describe how coherent the objects are on these attributes. The following definitions serve this purpose.Definition: Given two objects u, v $\in$ D and $\delta > 0$, we say that there exists a coherent pattern between u and v in subspace S, if formula (3) and (4) are true.

$$\forall i, j \in S, d_{ij} = \max(u_i - u_j) - \min(v_i - v_j) \leq \delta \quad (3)$$

$$\forall i, j \notin S, d_{ij} = \max(u_i - u_j) - \min(v_i - v_j) > \delta \quad (4)$$

To tell whether two objects in D exhibit a coherent pattern in a given subspace S, it is essential to describe how coherent the objects are on these attributes. The following definitions serve this purpose.Definition: Given two objects u, v ∈ D and δ > 0, we say that there exists a coherent pattern between u and v in subspace S, if formula (3) and (4) are true. Subspace S is defined by the set of bounded dimensions (or subspaces), in which objects u and v have a similar shifting pattern. That is to say, if the rank of the two objects on two arbitrary attributes in S is less than a user-specified threshold d, we say that the two objects have a coherent pattern, as illustrated in Fig. 2. The minimal variation of object v on attributes i, j is Δ, while the maximal variation of u is Δ + δ, and the difference is less than δ. If all pairs of attributes in S satisfy this, u, v have coherent pattern.

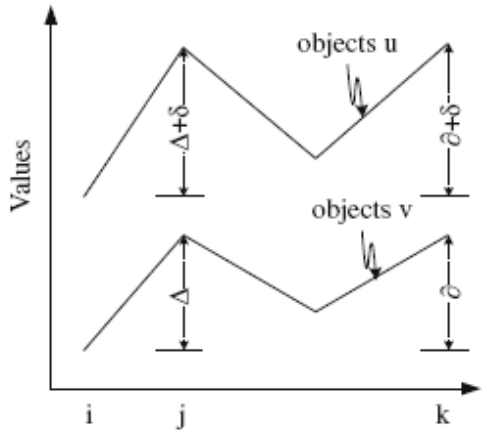


Figure (2): *A coherent pattern between two objects*

In this paper a novel approach is proposed for attribute selection inferred from the frequent pattern-based clustering algorithm.

3. Dimensionality Reduction:

Dimensionality reduction can be provided by feature selection and by feature extraction. Feature extraction method creates a subset of new features by combination of existing features, while feature selection method chooses a subset of all features that is more informative (more relevant to the target class). Both are utilized as a preprocessing stage for classification to improve its accuracy, reduce memory space and processing time required for classification and to reduce the cost of gathering data, noting that irrelevant features could be represented as a noisy feature that could decrease the accuracy of the classification process [13].

An example of feature extraction methods is Independent Component Analysis ICA and Principal Component Analysis PCA. ICA removes the redundancy in features by making them as much independent from each other as possible [14]. In PCA, the new variables are linear combination of the original features, chosen to capture as much as the original variance as possible [15].

On the other hand feature selection methods are divided into two types: univariate and multivariate feature selection methods. Univariate methods evaluate the relevance of features individually where it provides the discriminatory power (ability of the feature to discriminate between different classes) of the feature [13], each feature is considered individually at a time. An example of univariate methods is CHI Square method and Mutual information MI method [16] which measures the dependency between each feature f and the target class c Where f and c are independent if: P(f, c) = P(f) P(c). MI is calculated as follows in formula (5):

$$MI(X, Y) = \int P(X, Y) \log P(X, Y)\, dX\, dY \quad (5)$$

MI is used later in [13] in another form to improve the discrimination between confusable classes (enlarge the separation between the correct class and other competing class). Such that for a feature X, the discriminating information for class Ci versus class Cj is measured as follows in formula (6)

$$I_i(X) = \int P_i(X) \log P_i(X)/P_j(X)\, dX \quad (6)$$

Where Pi(X) and Pj(X) are the probability density functions of class Ci and class Cj for sample X. And the total averaged information for discriminating class Ci from class Cj

(referred to as divergence) is given by formula (7):

$$D_{ij}(X) = I_{ij}(X)+I_{ji}(X) \quad (7)$$

Multivariate methods consider a subset of features together. Some researches on classification techniques show the implementation of a statistical approach for creating a classifier and identifying a small number of relevant features simultaneously [17]. Other methods depend on the creation of the covariance matrix and corresponding eign values and eign vectors created in the PCA method [18]. Another example [19] is the correlation based feature selection which is a pair-wise feature selection method depends on the correlation coefficient between two features fi and fj using σij which is calculated as follows in formula (8):

$$\sigma_{ij} = COV(f_i , f_j) \; / \; \sqrt{VAR(f_i) * VAR(f_j)} \quad (8)$$

Feature selection method is mainly used an evaluation function to evaluate the important degree of feature to the target class. Usually, the assessed value is calculated first, then the feature which is assessed value is lower than setted threshold are removed. These methods are known as filter methods where they are independent on the predictor. In practice, people often use their experience to set an initial value, then debug threshold repeatedly according to the classification results [20]. This method is defined as wrapper feature selection method, which uses any machine learning algorithm as a black box, and search the space of all possible feature subsets to build a predictor with optimum performances. An example of Forward selection is [21] which uses Information gain to select best features then use KNN and SVM to evaluate the selected features. Wrapper methods have many types like sequential forward selection and sequential backward elimination. Forward selection starts with the single most informative feature and iteratively adds the next most informative features in a greedy fashion [22].

The approach proposed in this paper uses multivariate wrapper feature selection method.

4. Missing Data Techniques:

Statisticians have identified three classes of missing data. The easiest situation is when data are missing completely at random (MCAR) where the probability that a variable is missing is the same for every record. If the probability that a value is missing depends only on the value of other variables, we say that it is missing at random (MAR). If the missingness depends on the missing value, data are not missing at random (NMAR), where missingness of 'Y-variable' is conditional on some other 'X-variable' observed in the data set, and this is a problem for many statistical MDTs. This happens, for instance, when we collect data with a sensor which is not able to detect values over a particular threshold. The easiest way to obtain a complete data set from an incomplete one is to erase missing items. If we do not want to lose data and perhaps information, we may try to guess missing items. This process is generally called imputation, for example mean imputation and most common attribute value (mode) imputation, but they have been compared against other MDTs where they usually regarded as bad methods by statisticians, because the standard deviation of the sample is underestimated even when data are MCAR [23, 24].

The work in [25] also divides the missingness mechanisms according to the target to be predicted. Informative occurs when the fact that a value is missing provides information about a classification target. Non-informative occurs when the distribution of missing values is the same for all the classes of values.

The approach proposed in this paper uses the deletion of missing data approach when the data is not informative.

5. Problem formulation:

Problems of Data classification can be summarized in this paper into six issues: Dealing with continuous data, the curse of dimensionality, The Univariate / Multivariate feature selection problems, the dependence on threshold, the existence of noise and missing data and finally Rule Extraction difficulty.

Continuous Data

Some Classification techniques like decision trees tend to perform better when dealing with discrete/categorical features [4]. One of the difficult problems in classification is to handle quantitative data appropriately [26], thus, it is often necessary to transform a

continuous attribute into a categorical attributes using discretization process [2, 27]. Also Gini index which is used for classification (Decision tree classification) in attribute selection, biases multivalued attributes, also tends to favor tests that result in equal sized partitions and purity in all partitions, the technique in [28] tries to solve these problems by making a change in the attribute selection technique.

The curse of dimensionality

One of the important problems facing the development of a practically usable classifier is the high number of features (curse of dimensionality), where some of them may be irrelevant to the classification or redundant [29]. Feature selection methods is usually preferred for such problem rather than feature extraction methods like PCA, as it is problematic when there are a large number of irrelevant features that could mask the real classes (or clusters) [1]. Beside in PCA, the number of features transformed is predefined (user dependent) that could result in misclassification.

Multivariate feature selection problem

Univariate approaches are simple and fast, therefore appealing and popular. However, they assume that the features are independent [30]. Bayes belief models, as an example, will be computationally intractable unless an independence assumption (often not true) among features is imposed. One of the approaches who try to solve this problem is [31] where the proposed technique aim to extract patterns from the input models and combine them with the decision tree to give interpretable rules. Several limitations restrict the use of multivariate approaches. Firstly, they are prone to overtraining, especially in p>>n (many features and few samples) settings. Secondly, they may be computationally expensive, which prevents them from being applied to a large feature space [32, 33].

The dependence on threshold

The threshold is an important factor in many feature selection methods. However, the threshold is very difficult (non-trivial) to determine. An inappropriate threshold value may result in too many or too few patterns, with no coverage guarantees [10]. In theory, there is no good solution. The existing approaches depend on a debugging scope that is often too great to be easy to determine the threshold [20].

Missing Data

There exist many techniques to manage data with missing items, but no one is absolutely better than the others. Different situations require different solutions. As Allison says, “the only really good solution to the missing data problem is not to have

any [23]. The study in [35] perform analysis on the performance of ANN in case of missing data and replaces the missing value by the normal value, but it faces a limitation which is excluding the cases of missing more than one value.

Rule Extraction

Finally the rule extraction methods may be difficult in some classification methods like in neural network.

The technique proposed, depends on pattern based subspace clustering as way to solve some of the above problems.

6. The Proposed model: Pattern-based classification:

In this paper, the proposed model a new classification model to face five main challenges. 1-Solves the problem of dimensionality. 2- Handles continuous data in a better performance. 3- Finds the threshold value used pattern-based classification automatically. 4- Handles missing data partially. 5- Uses multivariate feature selection method to deal with feature-feature dependence.

The model is based on a frequent pattern-based subspace clustering technique. The selected data set is of a given and known class so we will use the definition of the Pattern-based subspace cluster but in a reverse order, as we are sure that there is a coherent pattern in this dataset (since they are in the same class) but we are going to use the definition to extract the set of attributes (subspace) S that includes the coherent pattern.

Briefly, the model works as follows: In the training phase of this approach we determine the frequent patterns appears for each class and determine the features (attributes) that produce such patterns according to different values of delta. Then in the testing phase we use these patterns for the selected features only to determine whether the tested object in the class corresponding to these patterns or not. According to the minimum error percentage value, the delta value and the used patterns are selected to be used for the classification of objects of the required class.

For example, in the training phase, patterns are evaluated and sorted as in fig 3 then the following patterns {[9, 15] [14, 15] [9, 14] [5, 14] } are selected (the first four patterns) to classify objects in class 1 , where each object in class 1 should have the delta values of attributes 9 and 15 is less than a certain value, where the delta δ equals to 29 in this example, and so the delta values of attributes 14 and 15, between 9 and 14 and between 5 and 14. In the testing phase we test the objects using 1 pattern, that the

difference between values of the attributes of 9 and 15 is δ, and find the error percentage, then make the test for the first 2 patterns and so on until we find the minimum error percentage

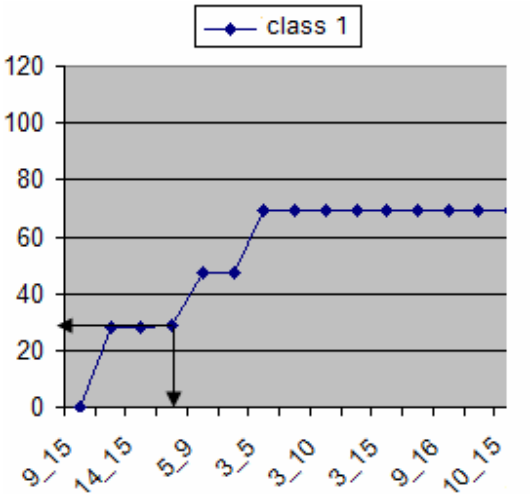


Figure (3): *Selection of some patterns (four) with the least delta values*

The steps of the proposed model in the following pseudo code in figure, fig 4:

Algorithm of the Proposed Module

- int NumOfTraining //number of objects in the training set of class x.
- int NumOfTesting //number of objects in the testing set of objects in different classes including class x.
- int NumOfAttribute //number of attributes (features)
- float[] Delta //array of D_{ij} (delta value for Pattern ij) in the training class
- float minErrPercent = 0.0 //Minimum error percentage
- float Final_Delta //Final output Delta used for classification
- int[][] Pattern_ij // Array of ij patterns used for classification
  // The array is initially empty.
- int[][] Final_ij //Final ij patterns used for classification, initially empty

- For i:1 to NumOfAttribute
  - For j: i+1 to NumOfAttribute
    - $D_{ij} = \max((u_i-u_j)-(v_i-v_j))$ For every two distinct objects
    - If (u_i-u_j) and (v_i-v_j) have the same sign (Same trend) for every two distinct objects in the training set of class x.
      - Add D_{ij} to Delta[]
- Sort the Delta[] in an ascending order.
- For k:1 to Delta[].size()

- Delta = Delta[k]
- Add The pattern ij of D_{ij}=Delta[k] to the array Pattern_ij
- For every two distinct objects (u and v) in the testing array
  - If Delta < $((u_i-u_j)-(v_i-v_j))$ where ij is every pattern in Pattern_ij
    - The object u and v is of class x
- Compare the classified objects with the real objects in the training set to find the error percentage,
- If the error percentage < minErrPercent
  - minErrPercent = error percentage
  - Final_Delta = Delta
  - Final_ij = Pattern_ij

- Return Final_Delta and Final_ij for classification of objects of class x

The steps of the proposed model in details:

Use the training dataset to evaluate the patterns that exist in class 1. In the frequent Pattern-based subspace cluster, the set of objects is unknown and the set of attributes is given while in this definition the set of attributes is unknown and the set of objects is given [3, 36].

So the definition will be changed as follows:

Given two objects u and v in D, and these two objects are in a class 1, we have the attributes i, j is in the space S, dij is calculated as in formula (9).

$$\forall u, v \in C, d_{ij} = abs((u_i - u_j) - (v_i - v_j)) \quad (9)$$

If the trend of change from ui to uj is opposite to that from vi to vj, for example if ui > uj while vi < vj then this ij pattern is excluded by setting the dij by -1.

Find the maximum dij values in a matrix of size ixj calculated in step one where there are different dij for each pair of objects in the same class. Then sort these values as in figure 4 in a single dimension array after removing the dij values of -1.

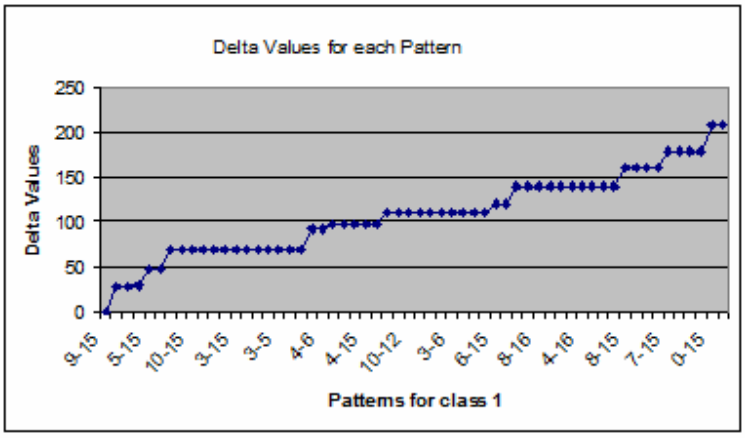


Figure (4): *Sorted Delta values for all patterns (58 pattern out of 164) after removing patterns of -1*

Then the model starts by testing the testing set objects, that contains objects related and not related to the needed class, using the first smallest pattern ij, where delta δ = dij equals to its value in the sorted array. We use the original definition of the pattern-based subspace cluster to classify other testing objects by proving the correctness of the definition in formula 7 and 8.

Given two objects u, v in D and δ> 0, we say that there exists a coherent pattern between u and v in subspace S, if the formulas (10) and (11) are true:

$$\forall i, j \in S, d_{ij} = \max(u_i - u_j) - \min(v_i - v_j) \leq \delta \quad (10)$$

$$\forall i, j \notin S, d_{ij} = \max(u_i - u_j) - \min(v_i - v_j) > \delta \quad (11)$$

The model repeats the test for 2 patterns using the delta dij of the second pattern and so on for 3 patterns using the delta dij of the third pattern and the method is repeated until the number of patterns used reaches the size of the sorted array or the error percentage is zero. Then returns back the (list of pattern) and the delta used that shows the minimum error percentage and minimum number of patterns. Note that after this step the used threshold δ is not user defined any more as in figure 5, the least number of patterns of

minimum error percentage are selected.

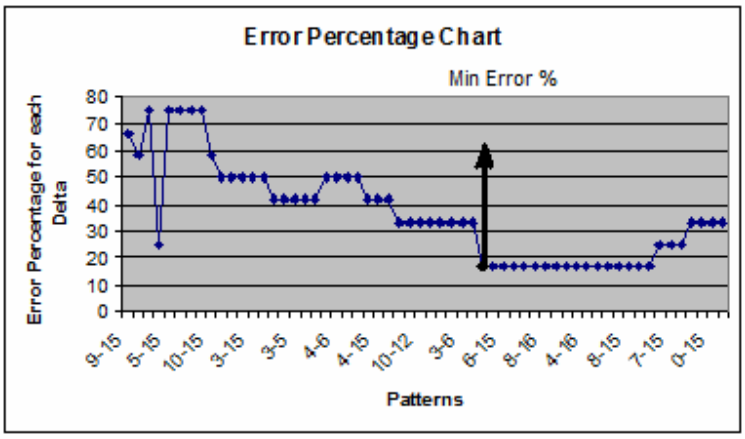


Figure (5): *Error percentage after using delta values of each pattern*

Finally if the accuracy of the results is good enough we can use it for the rest of the dataset. If the rank of two arbitrary objects on two attributes is less than a threshold δ, we say that the two attributes are in the set of attributes (Subspace) S.

Later on we could use the testing objects and the training objects together, so that the coherent between both the known and unknown could show the class of the tested objects.

7. Experiments and Results:

We apply such model within bioinformatics (Gene classification) and medical data set (Thrombosis disease classification).

Gene Classification

Data mining methods have been widely applied in bioinformatics to analyze gene functions, gene regulations, cellular processes, and subtypes of cells. Gene Classification is one important issue in gene expression data analysis because it is a basis for prediction of the function of unknown genes and much work on gene microarray analysis indicates that high correlation may exist between gene expression patterns and diseases patterns [37]. In fact, the expression levels of two closely related genes may rise and fall synchronously in response to some environmental stimuli. Unfortunately, conventional distance functions can not model this similarity effectively

since the expression levels may not be close in most cases. Thus it is natural to apply the pattern-based clustering analysis to microarray data [36].

We perform an experimental study on the efficiency and effectiveness of our model with a yeast gene expression matrix with 17 conditions (attributes or features) and 40 genes which is a part of a bigger matrix [38]. We know previously 22 genes that have the similar characteristics, ten of these set is going to be used as a training data and 12 in 31 objects which are the testing set.

The results appears from our model is

Minimum Error Percentage is 16.666666666666657% for the following patterns of a delta = 110.0:

[9, 15] [14, 15] [9, 14] [5, 14] [5, 15] [5, 9] [15, 16] [14, 16] [10, 15] [10, 14] [9, 16] [9, 10] [3, 15] [3, 14] [3, 10] [3, 9] [3, 5] [4, 14] [4, 5] [6, 14] [4, 6] [6, 10] [5, 16] [5, 10] [4, 15] [4, 9] [13, 15] [11, 15] [10, 12] [9, 13] [9, 11] [4, 11] [3, 6] [3, 4] [2, 15]

Correlated objects in class 1 are:

0, 1, 3, 5, 8, 9, 2, 4, 6, 23, 10, 7

The classification time is 0.02 seconds 35 patterns out of 161 patterns appear to be used to find out correlated objects as in figure 4 and figure 5 to the required class with error percentage is 16,6% where the delta evaluated and used is 110.0.

Thrombosis disease classification

On the other hand, we test our model on a database collected at Chiba University hospital from the outpatient clinic of the hospital on collagen diseases (are auto-immune diseases). A thrombosis is one of the most important and severs complications in collagen diseases. It is important to detect and predict the possibilities of its occurrence. Domain experts are very much interested in discovering regularities behind patients' observations [39]. Thrombosis has four main levels or degrees, which 0 (negative or no thrombosis), 1 (positive and the most severe one), 2(positive and sever) and 3(positive and mild).

We perform an experimental study on 2 sets only of data which are of 0 and 1 degrees of thrombosis. We worked on 12 cases of 16 attributes (tests) of patients of thrombosis of degree 1. Then we test resulted patterns and delta value on 5 cases where the first 3 cases have thrombosis of level one.

The results appear as follows:

Minimum Error Percentage is 33.33333333333334% for the following patterns of a delta = 7.3

[9, 12] [8, 12] [5, 9] [5, 8] [4, 5] [4, 9] [4, 12] [4, 8] [6, 13] [12, 13] [6, 8] [8, 13] [6, 9] [9, 13] [5, 13] [4, 13] [5, 7]

```
Correlated objects in class 1 are:
0, 2
The classification time is 0.02 seconds
```

17 patterns appear to be used to find out correlated objects as in figure 6 and figure 7 to the required class with error percentage is 33.3% where the delta evaluated and used is 7.3.

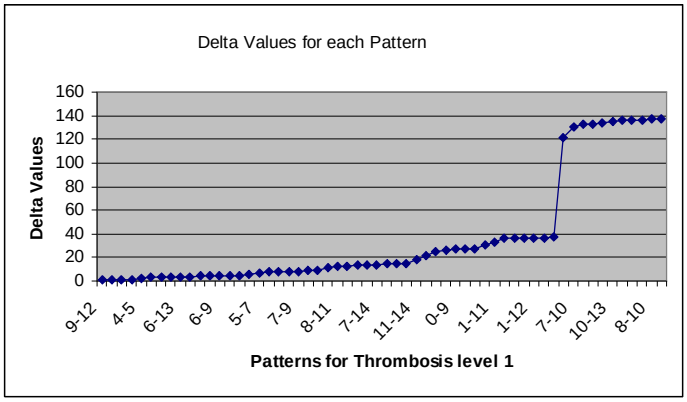


Figure (6): *Sorted array The Delta values of Thrombosis of level 1*

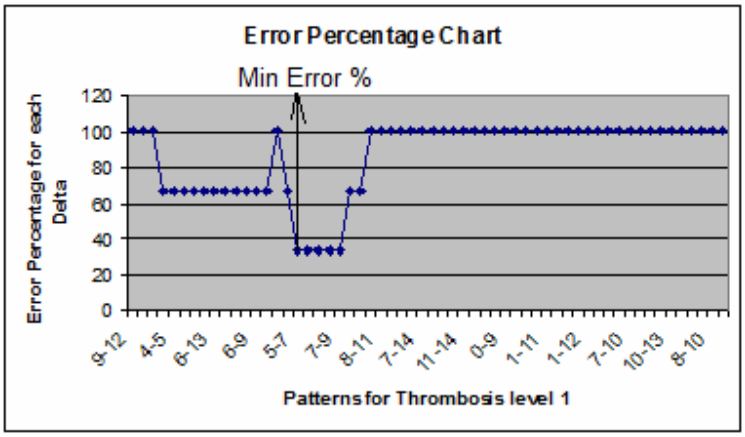


Figure (7): *Error percentage after using delta values of each pattern and selecting pattern of the minimum error percentage*

Images of hand-segmented classification

Another application on this model is a continuous data collected from University of Massachusetts [40]. The subset of segmentation data is a randomly drawn instance from a database of seven outdoor images. The images were hand-segmented to create a classification for every pixel. The number of training data is 40 instances each of 19 continuous attributes. We classify only two classes out five defined classes in this data which are grass and path classes, the results appears to be of 0 % error percentage and the classification time is 0.04 seconds.

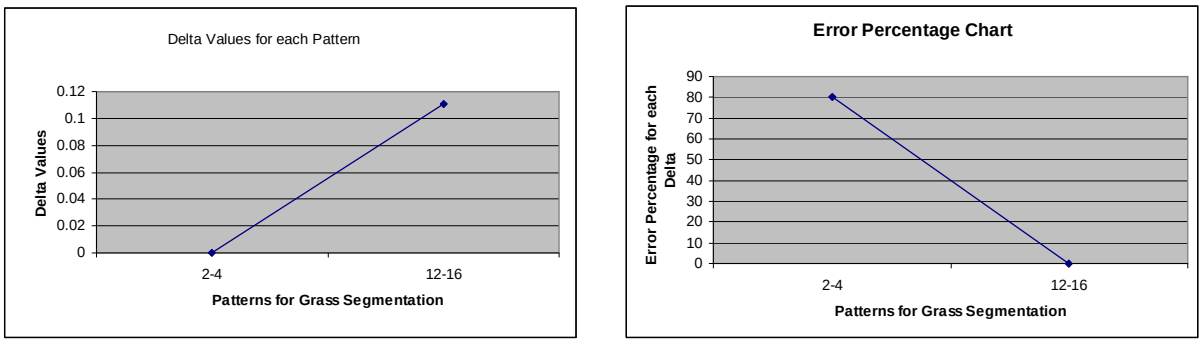


Figure (8): *only 2 selected patterns are selected since the error percentage is zero*

A Comparison with other classification models:

The goal of any classification model is to generate more certain, precise and accurate system result in a good performance. A comparison on the accuracy of prediction and the time of classification should take place with other models to prove the strength of our model. We use six different models which Bayesian Network (BN), Naïve Bayesian Network, Decision table, BFTree, IB1 and classification via clustering models. Bayesian Networks is a graphical model for probability relationships among a set of variable features. Decision trees are trees that classify instances by sorting them based on feature values, a BFTree is type of decision tree that uses a best first method of determining its branches (shi-2007) [27]. A lazy model-based algorithm focus its effort on classifying the particular event in question, and would also return a rationale that may help the person interpret the validity of the prediction, IB1 (which uses the nearest instance for final prediction) is used because it represents the core and simplest lazy learning method [41]. The following table, table 1 contains a comparison according to the error percentage and time of learning phase of each model.

***Table (1)**: Comparison between different classification model according to error percentage and time taken*

	Our Model	Classification Via Clustering	IB1	BFTree	Decision Table	Naïve Bayes	Bayes Network
Gene							
Error %	16.6%	92%	66%	100%	66%	75%	75%
Time	0.012	0.01	0.0 s	0.01 s	0.04 s	0.0 s	0.01 s
Thrombosis							
Error %	33%	0%	100 %	80%	80%	100%	80%
Time	0.02 s	0.01 s	0.0 s	0.03 s	0.02 s	0.0 s	0.0 s
Segmentation							
Error %	0%	0%	0%	0%	0%	0%	0%
Time	0.04 s	0.06 s	0.0 s	0.07 s	0.11 s	0.0 s	0.03 s

8. Contribution:

Beside classification accuracy and time complexity of each model the six characteristics or requirements maintained above should be considered for any classification technique. Our model tries to support most of these requirements.

Dealing with continuous data: The approach deals with continuous data without the need of discritization.

The curse of dimensionality: The approach uses the concept used in pattern based clustering algorithm that enables it to determine the patterns that are hidden between the features of each class to remove features that are irrelevant and noisy.

The model uses a Multivariate feature selection method where the idea that distinguishes it from other Multivariate approaches that it depends on the trend of variation of values among different features of object which is required in real life cases. One of the approaches that tries to apply this is PCA as it uses the covariance to measure the variation of pair of features (I, j) together as appears in formula (12).

$$\sigma_{ij}^2 = \mathrm{cov}(i,j) = \frac{1}{M}\sum_{m=1}^{M}(x_{im} - \mu_i) * (x_{jm} - \mu_j)$$

Where M is the number of objects in training space and μi and μj is the means of the features i and j respectively. The problem of the usage of covariance is that it depends on the deference between each feature value and the mean value of the feature for all objects. In real life cases, this may not be applicable, as in some cases, a feature value that differs from the mean or mode value could be normal. For example in medical life some people have their blood pressure higher than the normal case while they are

healthy and suffers from no disease.
The independence on threshold where the approach is a wrapper method that is a forward feature selection method. It solves a problem that appears in the pattern based clustering model and all the classification models that use the clustering algorithms which is the dependence on the user in determining the threshold value for selecting the feature. It is built on a hypothesis that if features are sorted according its relevance to the a certain class (with putting into consideration the correlation between feature) is fully correct, and classification is performed using the first attribute, then performed using the first two attributes, and so on until you select all the attributes. Then the error in classification should be decreasing until a certain point (threshold) then start to decrease. This could appear clearly in the following figure (9).

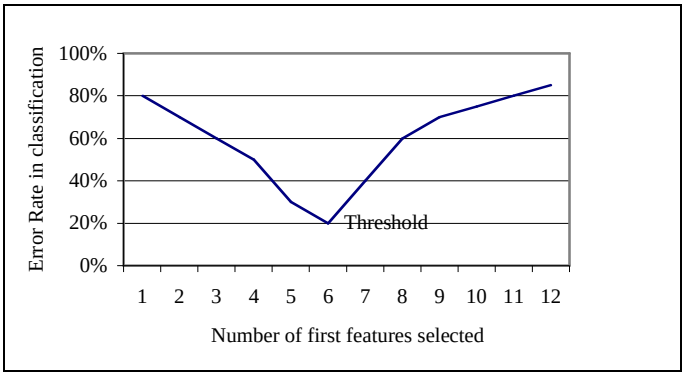


Figure (9): *The change of the error rate in classification according to the number of features selected*

This is proved experimentally in this paper where it is clear in figure (5) and figure (6) when applying the proposed model on Gene dataset and thrombosis dataset respectively.

The existence of noise and missing data. The model should use a complete data set in training while in testing the classification depends only on the features selected (informative features). If the missing data is not informative (not of the selected features), then it won't affect the classification and it will be simply ignored. If the missing data is informative the object will be deleted, but we are going to make an imputation of this case in future work.
Finally Rule Extraction difficulty, which is considered also as a future work.

There is an extremely large literature on classification learning, including the use of clustering to augment classification. But most of these methods have a deficiency that they fail to handle all of these features altogether. For example methods that use

clustering algorithms in classification like K-mean clustering and DBSCAN do not put into consideration the curse of dimensionality [42].

9. Conclusion And Future Work:

There are various approaches to determine the performance of classifiers. The performance can most simply be measured by counting the proportion of correctly predicted examples in an unseen test dataset and the time of the learning and testing methods of each model. Although or perhaps because many methods of ensemble creation have been proposed, there is as yet no clear picture of which method is better. The proposed approach handles most of the discussed requirements of data classification techniques in a single model. Some models uses the clustering algorithm like pattern based clustering model to select a part of the features in high dimension data, then uses the selected features in a known classification model like decision trees and Bayesian networks. However the proposed model uses a concept in the clustering model itself to make the classification of the data. A comparison is made with six different classification models according to the accuracy of performance and the time of learning. Our model proves its efficiency and its competency according these criteria.
Missing or incomplete data is a usual drawback in many real-world applications of pattern classification. Our planned future work is to find an appropriate missing data imputation for informative data in our model and to find a way for rule extraction.

References:

[1] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, “Introduction to Data Mining", Addison-Wesley, May 2005. (ISBN:0-321-32136-7).

[2] Nikolaos Mastrogiannis, Basilis Boutsinas, Ioannis Giannikos, “A method for improving the accuracy of data mining classification algorithms”, Computers & Operations Research, Volume 36, Issue 10, October 2009, pp. 2829-2839, ISSN 0305-0548.

[3] Jiawei Han, Micheline Kamber, ”Data Mining: Concepts and Techniques”, Second Edition, 2006.

[4] Thair Nu phyu, “Survey of Classification techniques in Data Mining”, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, March 18 - 20, 2009, Hong Kong.

[5] Yun-Chao Gong, Chuan-Liang Chen, "Semi-Supervised Method for Gene Expression Data Classification with Gaussian and Harmonic Functions", Pattern Recognition, 2008. ICPR 2008. 19th International Conference, 2008, pp. 1-4.

[6] Christy, A. and P. Thambidurai, "Feature Selection for Efficient Text Categorization and Knowledge Discovery Using Classification Techniques", Asian Journal of Information Technology 5(8), 2006, pp. 872-876.

[7] Roland Nilsson, Jose M Pena, Johan Björkegren and Jesper Tegnér, "Consistent feature selection for pattern recognition in polynomial time", Journal of machine learning research, 2007, (8), pp. 589-612.

[8] Shoushan Li, Rui Xia, Chengqing Zong, Chu-Ren Huang, "A Framework of Feature Selection Methods for Text Categorization", Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, 2-7 August 2009, pp. 692–700.

[9] R. Xu, D. Wunsch, "Survey of clustering algorithms", IEEE Transactions on Neural Networks 16 (3), 2005, pp. 645–678.

[10] Hassan M. Malik and John R. Kender,"Classification by Pattern-based Hierarchical Clustering", In From Local Patterns to Global Models Workshop, ECML/PKDD, 2008

[11] T. Padma, Madhavi Latha, K. Jayakumar, "Decision Making Algorithm through LVQ Neural Network for ECG Arrhythmias", ICBME 2008, Proceedings 23, pp. 85-88.

[12] Abderrahmane Boubezoul, Sebastien Paris, Mustapha Ouladsine, "Application of the cross entropy method to the GLVQ algorithm", Pattern Recognition 41, 2008, pp. 3173-3178.

[13] Jingbo Zhu, Huizhen Wang, Xijuan Zhang,"Discrimination-Based Feature Selection for Multinomial Naïve Bayes Text Classification", ICCPOL, 2006, pp. 149-156.

[14] Nojun Kwak, Chong-Ho, Jing Young Choi,"Feature Extraction Using ICA", ICANN, 2001, pp. 568-573.

[15] Z. Cataltepe, H. M. Genc, T. Pearson,"A PCA/ICA Based Feature Selection Method and its Application for Corn Fungi Detection", Eusipco (European Signal Processing Conference), Poland, 2007.

[16] Saeys Y, Inza I, Larranaga P,"A review of feature selection techniques in bioinformatics", Bioinformatics, Vol. 23, No. 19. , 2007, pp. 2507-2517.

[17] L. R. Grate, C. Bhattacharyya, M. I. Jordan and I. S. Mian,,"Simultaneous Relevant Feature Identification and Classification in High-Dimensional Spaces", Proceedings of the Second International Workshop on Algorithms in Bioinformatics, 2002, pp. 1-9.

[18] Alexey Tsymbal , Seppo Puuronen , Mykola Pechenizkiy , Matthias Baumgarten , David Patterson,"Eigenvector-Based Feature Extraction for Classification", Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, 2002, pp. 354 - 358

[19] KRZYSZTOF MICHALAK, HALINA KWA´SNICKA,"CORRELATION–BASED FEATURE SELECTION STRATEGY IN CLASSIFICATION PROBLEMS", Int. J. Appl. Math. Comput. Science, 2006, Vol. 16, No. 4, pp. 503–511.

[20] yanling li, Li Song,"Threshold determining method for feature selection", Proceedings of the 2009 Second International Symposium on Electronic Commerce and Security, 2009, Volume 02, pp. 273-277.

[21] Jinjie Huang, Yunze Cai, Xiaoming Xu,"A hybrid genetic algorithm for feature selection wrapper based on mutual information", Pattern Recognition Letters, 2007, Volume 28 , Issue 13, pp. 1825-1844.

[22] Carmen Lai, Marcel J. T. Reinders, Lodewyk F. A. Wessels, "Random subspace method for multivariate feature selection", Pattern Recognition Letters (PRL), 2006 27(10, pp. 1067-1076

[23] Magnani M.,"Techniques for Dealing with Missing Data in Knowledge Discovery Tasks", department of computer. Science, University of Bologna, 2004

[24] P. D. Allison,"Missing data. Sage Publications", Inc, 2001.

[25] Pedro J. Garcıa-Laencina, Jose-Luis Sancho-Gomez, Anıbal R. Figueiras-Vidal, "Pattern classification with missing data: a review, Neural Computing and Applications", 2009

[26] B. Chandra, P.Paul Varghese, "Fuzzifying Gini Index based decision trees", Expert Systems with Applications 36, 2009, pp. 8549-8559

[27] Gagne II, David John; McGovern, Amy and Brotzge, Jerry, "Using Multiple Machine Learning Techniques to Improve the Classification of a Storm Set", Preprints of the Sixth Conference on Artificial Intelligence and its Applications to the Environmental Sciences, 2008.

[28] Quoc-Nam Train,"Mining Medical Databases with Modified Gini Index Classification", Proceedings of the Fifth International Conference on Information Technology: New Generations, 2008, pp. 195-200.

[29] C. Shang and Q. Shen,"Aiding classification of gene expression data with feature selection: a comparative study". Computational Intelligence Research, 2006, 1(1), pp. 68-76

[30] S. Gunal and R. Edizkan,"Subspace based feature selection for pattern recognition", Information Sciences, 2008, 178(19), pp. 3716–3726.

[31] Pierre Geurts, "Pattern Extraction for Time Series Classification", Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, 2001, pp. 115 - 127.

[32] Carmen Lai, Marcel J. T. Reinders, Lodewyk F. A. Wessels,"Random subspace method for multivariate feature selection", Pattern Recognition Letters (PRL) 27(10, pp.1067-1076, 2006

[33] Yiming Yang, Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", 1997, ICML, pp. 412-420.

[34] Pedro J. Garcıa-Laencina, Jose-Luis Sancho-Gomez, Anıbal R. Figueiras-Vidal, "Pattern classification with missing data: a review", Neural Computing and Applications, 2009

[35] Jihong Guan, Yanglan Gan, Hao Wang,"Discovering pattern-based subspace clusters by pattern tree", Knowledge-Based Systems 22 (2009), pp. 569-579.

[36] X. Zhang, W. Wang,"Mining coherent patterns from heterogeneous microarray", in: Proceedings of the 15 th ACM International Conference on Information and Knowledge Management, 2006.

[37] X. Zhang, W. Wang,"Mining coherent patterns from heterogeneous microarray", in: Proceedings of the 15 th ACM International Conference on Information and Knowledge Management, 2006.

[38] Chiba University hospital DataBase "http://lisp.vse.cz/pkdd99/"

[39] Segmentation data from Weka Project " http://www.cs.waikato.ac.nz/~ml/weka/"

[40] Kai Ming Ting, "Discretisation in Lazy Learning Algorithms", Artificial Intelligence Review, Feb 1997, Volume 11, pp. 157-174.

[41] Jeffrey Erman, Martin Arlitt, Anirban Mahanti, "Traffic Classification Using Clustering Algorithms", SIGCOMM'06 Workshops, Pisa, Italy, September 11-15, 2006, pp. 281-286.