Fetal State Classification from Cardiotocography Based on Feature Extraction Using Hybrid K-Means and Support Vector Machine

Nurul Chamidah, Ito Wasito
Faculty of Computer Science
University of Indonesia
Depok, Indonesia
nurulchamidah.2007@gmail.com, ito.wasito@cs.ui.ac.id

***Abstract*—Cardiotocography (CTG) records fetal heart rate (FHR) signal and intra uterine pressure (IUP) simultaneously. CTG are widely used for diagnosing and evaluates pregnancy and fetus condition until before delivery. The high dimension of CTG data are the problem for classification computation, by extracting feature we can get the useful information from CTG data, and in this research, K-Means Algorithm are used. After extracting useful information, data are trained by using Support Vector Machine (SVM) to obtain classifier for classifying the new incoming CTG data. Based on 10 cross validation, this method have a good accuracy to 90.64% using Cardiotocography Dataset obtained from UCI Machine Learning Repository. Data are classified into fetal state normal, suspicious, or pathologic class based on seven abstract features that extracted from twenty one original features and then trained using hybrid K-SVM Algorithm. This research shows the ability and capability of Hybrid K-SVM for classifying CTG dataset. In general, the experimental result of hybrid K-SVM show the better classification compare to SVM.**

Keywords—cardiotocography (CTG); fetal state; feature extraction; clasification; SVM; K-Means

I. INTRODUCTION

Healthy pregnancy, normal delivery and a healthy baby is the desire of almost mothers. These conditions are supported with regular prenatal care. For the mother, the examination is useful to detect problem of pregnancy, preparing mentally and physically, know the condition of pregnancy, and to determine the appropriate delivery method based on the results of the examination. For baby, it can be seen by examination of the baby's condition so that it can be maintained and minimize health risks at birth.

Cardiotocography (CTG) is one way of prenatal examination. CTG is an electronic method for monitoring the condition of the fetus during pregnancy and childbirth or within. CTG recorded signals from the baby's heart rate or fetal heart rate (FHR) and uterine contractions or intrauterine pressure (IUP) at the same time.

The use of CTG reduces the incidence of birth asphyxia, where the baby was born without a heartbeat caused by hypoxia. However, this method has side effect on the increase in Caesarean section births [1]. The use of CTG is highly dependent on the readings of FHR pattern, where the reading of this pattern shows a low standard in clinical practice even there is the Guidelines for CTG readings is exist. The variation of this interpretation is done not only by inter-observer but also intra – observer, the consequences of this various CTG interpretation is rising birth to Caesarian section and lack specify in detecting acidosis. Because manual interpretation rises an error, then computing and data mining techniques are necessary to analyze and classify the data CTG to avoid human error.

The challenge in the medical domain is to extract knowledge from data such as CTG diagnosis. Although at this time, there is no consensus the best methodology as the baseline for estimating in CTG analysis [2], Neural network to classify the CTG data into 3 classes, normal, suspicious and pathologic have been conducted in [3]and [4]. The results of their study showed that using ANN for both normal and pathologic class data shows a good result, and weak for suspicious class. The comparison of the FCM method, K-Means, and ANN showed that ANN is the best method, but suspicious class cannot be properly classified and only achieve a precision of 0.58[5].

The high input dimensions of data are a problem in classification. If the dimension is high, then the training process of machine learning will require a large of CPU time. But there is no guarantee that the higher dimension and larger computational time is also going to improve the accuracy. With the redundant data and noise, increasing in accuracy is still a question and may even degrade accuracy. Therefore we need a method to solve this accuracy and computational time problem.

Study to reduce the feature of CTG by eliminating some features of the CTG to speed up the computation time [6] gives the highest sensitivity result is achieved at the Meta Selection with 75%. This feature elimination means removing information from the feature as a whole, in the sense that this method removes information to speed up the computation time. Prasad et. al. [7] looking for the best subset of features of breast cancer for use in the SVM training and provides result

 978-1-5090-0363-1/15/$31.00

in increasing accuracy and computational time because the reduced of feature dimensions. K-Means algorithm clusters features or characteristic of the data with unsupervised learning to obtain the important feature for use in the training data. Zheng et al [8] reduce the feature dimension in the case of breast cancer without removing information by applying feature selection. This study uses the K-Means algorithm to detect hidden patterns and SVM as classifier algorithm. Zheng's research show significant gains in achieving accuracy to 97.38% and speed up the computation time in training SVM, with feature reduces from thirty to six features. From the studies described above, this study will use SVM to classify the CTG data and K-Means algorithm to reduce the features.

II. METHODOLOGY

A. Data

The data used in this study are cardiotocography dataset from UCI Machine Learning in https://archive.ics.uci.edu/ml/. Dataset consists of measurements of fetal heart rate feature (FHR) and uterine contraction (UC) cardiotocograms classified by the obstetrician. 2,126 records with 23 attributes, where 21 is a feature and 2 types of classes, the class pattern (1-10) and fetal state class (N = Normal S = suspect, P = pathologic). In this study we will use fetal state class.

B. Feature Extraction

Proposed research method is shown in Fig. 1. The objective of this research is to classify new fetal state that hybridizes K-Means and SVM. To reduce dimension of feature, every class of CTG is separated to get abstract feature pattern of each class.

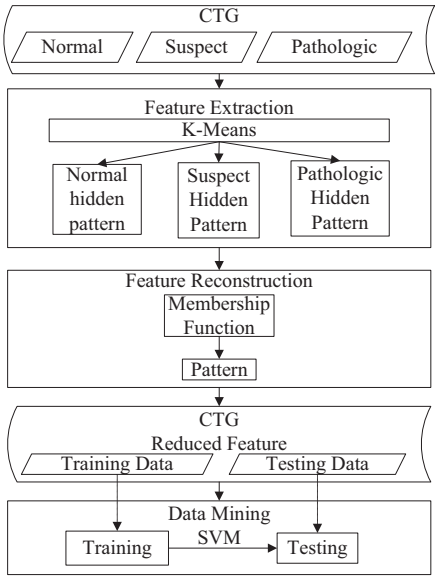

Fig. 1. Research Method

Feature extraction and selection is performed to obtain the hidden pattern of normal, suspect and pathologic fetal state separately with the K-Means algorithm. K-Means clusters the features that most affect fetal state by finding the shortest distance between the center (centroid) cluster and its members. Cluster membership is defined as follow:

$$\min_{\bar{X}_1,\bar{X}_2,\ldots\bar{X}_{nK}} \sum_{nk=1}^{nK} \sum_{i\in H_{nk}} \left\| X^i - \bar{x}_{nk} \right\|^2 \quad (1)$$

Where:

nK : Cluster index

H_{nk} : nK^{th} cluster set

$\bar{x}_{nk}$: Cluster's center or centroid point in cluster H_{nk}.

K-Means iteratively calculate the shortest Euclidean distance by adapting the centroid point location. The number of clusters is determined by calculating the Calinski-Harabasz criteria [9]. Calinski-Harabasz index is defined as:

$$CH_C = \frac{SS_B}{SS_W} \times \frac{(N-nK)}{(nK-1)} \quad (2)$$

$$SS_B = \sum_{x_{nk}\in C} |x_{nk}| \, \|\bar{x}_{nk} - \bar{X}\|^2 \quad (3)$$

$$SS_W = \sum_{x_{nk}\in C} \sum_{x_i \in x_{nk}} \|x_i - \bar{x}_{nk}\|^2 \quad (4)$$

Where:

SS_W : Distance between points in the cluster to its centroid (within cluster distance).

SS_B : Distance between centroid clusters to global centroid (between cluster distances)

$|x_{nk}|$: The number of nk^{th} cluster-

x_{nk} : nk^{Th} cluster

$\bar{x}_{nk}$: nk^{Th} centroid cluster

$\bar{X}$: Global mean of dataset

$C = \{x_1, x_2, .., x_{nK},\}$: clustering data X with the number of data N objects, to nK group and (group > 1)

Optimum cluster obtained from maximum CH index value. In this research, the number of optimum cluster is calculated in the range of cluster ($2 \leq nK \leq 21$) for each fetal state class.

C. Feature Reconstruction

Pattern of CTG is represented by cluster specifically. Symbolic CTG symbolized by the cluster center of the cluster. Symbolic CTG formed into normal, suspect, and pathologic datasets, and then its similarity is calculated with untested CTG. This process is useful to determine the suitability of the data CTG with a new pattern that has been found. To compute this similarity, fuzzy membership function is used [8] from the point of untested CTG to a pattern that has been identified.

$$fuzzy_{np}(X_j^i) = \begin{cases} 1 - \dfrac{\left|X_j^{\bar{x}np} - X_j^i\right|}{\max\left|X_j^{\bar{x}np} - X_j^n\right|} \; if \; (\min(X_j^n) \leq X_j^i \leq \max(X_j^n), \forall n \in H_{np} \\ 0, \; otherwise; \end{cases} \quad (5)$$

$$Pat_{np} = \frac{1}{D}\sum_{j=1}^{D} fuzzy_{np}(X_j^i), 1 \leq np \leq nK^{nor} + nK^{sus} + nK^{patho} \quad (6)$$

Where:

np : Index of new pattern

 978-1-5090-0363-1/15/$31.00

X_j^i : j[th] feature of i[th] original input
$X_j^{\bar{x}_{np}}$: j[th] feature of centroid $\bar{x}_{np}$ in cluster H_{np}
nK^{nor} : The number of pattern of normal fetal state
nK^{sus} : The number of pattern of suspect fetal state
nK^{patho} : The number of pattern of pathologic fetal state

Pattern obtained from the K-means clustering is an abstract CTG feature that different from the feature before [8]. Feature dimensions have been reduced with new abstract features that contain information form original features to each class. Then these new abstract features are used for training SVM classifier to obtain a normal class, suspect, and pathologic.

D. SVM

SVM gives a good accuracy (based on previous studies), so in this study we use SVM with polynomial kernel function.

$$maximize_\beta \left[\sum_{i=1}^{n} \beta_i - \frac{1}{2}\sum_{i,j=1}^{n} \beta_i \beta_j y_i y_j K(x_i x_j)\right] \quad (7)$$

$$subject\ to\ \sum_{i=1}^{n} \beta_i y_j = 0, 0 \le \forall \beta_i \le L \quad (8)$$

Where:
x : Training vector
y : Training vector label
β : Vector parameter of hyperplane *classifier*
K: Kernel function, calculating distance between training vector x_i and x_j
L: Penalty parameter to control the number of misclassified, the larger L, more accurate the classification result.

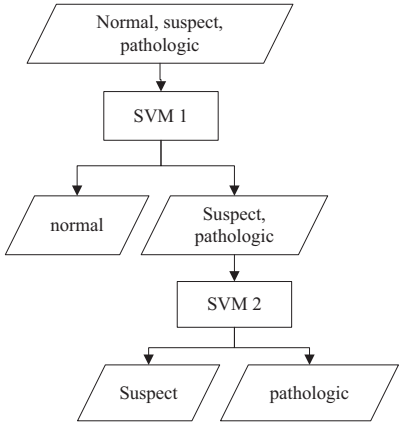

Fig. 2. Classification Model of SVM in Three classes

SVM is a classification algorithm that separates the two classes. In this study, to classify the CTG into three classes, we use SVM model that classifies normal and abnormal (suspect and pathologic) class. After normal class classifier is obtained, training for suspect and pathologic class to get the suspect and pathologic classifier is conducted. Fig. 2 shows the SVM models that were used in this study to separate the normal, suspect, and pathologic classes.

Training procedures in the model are summarized as follows:

- Step 1. Training data is divided into two classes, namely normal and abnormal (suspect and pathologic)
- Step 2. Conduct Training SVM 1 of the data in Step 1.
- Step 3. Training SVM 2 using suspect and pathologic data to obtain a classifier suspect or pathologic

III. Experimental Result

A. Feature Extraction

CTG dataset is separated by class of fetal state that previously had been performed normalization. Each of these data is clustered using K-Means, the number of clusters is obtained by evaluating the criteria Calinski-Harabasz. Evaluation criterion for the number of clusters with the cluster is done in the range of 2 to 21. Fig. 2 below shows the optimum number of clusters for each class of normal, suspect, and pathologic fetal state class of CTG data.

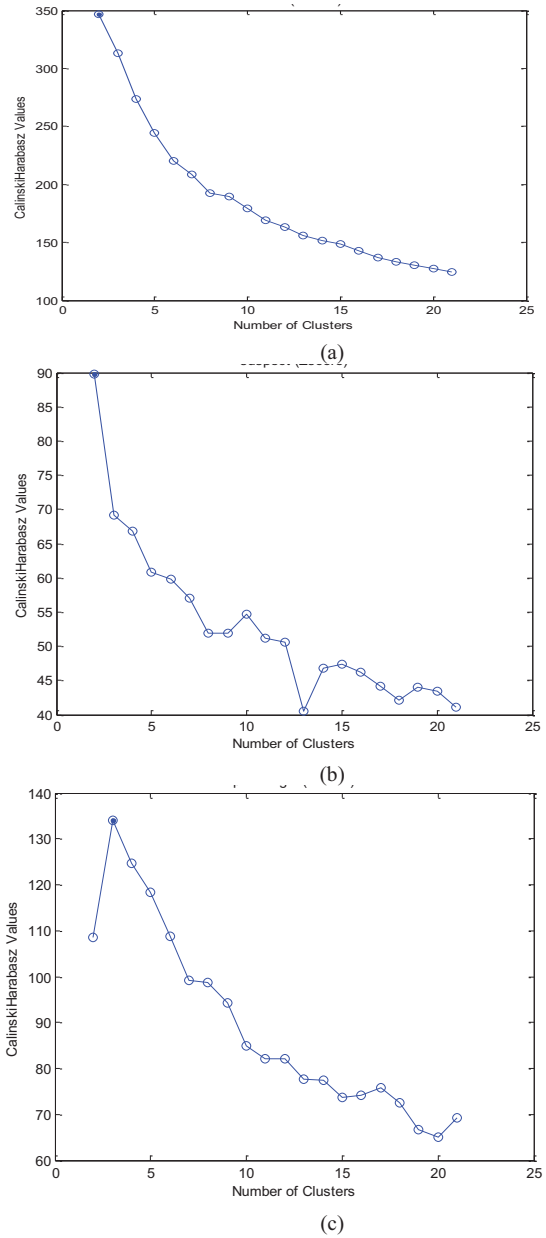

Fig. 3. Optimum number of clusters (a). Normal class, (b). Suspect class, and (c). Pathologic class

 978-1-5090-0363-1/15/$31.00 ©2015 IEEE

Fig. 3 shows Calinski-Harabasz index for normal, suspect, and pathologic class. Calinski-Harabasz criterion that indicates the optimum number of clusters is the criterion with the highest value. In Fig. 3 (a), the data of normal fetal state can be better when separated into two clusters as indicated by the value of the maximum Calinski-Harabasz. The separation of the fetal state suspect class is shown in Fig. 3 (b) that visualize the optimum number of clusters with two clusters, and pathologic fetal state class is well-separated according to the criteria Calinski-Harabasz at three clusters that have been shown in Fig. 3 (c), where the top of the chart are at three number of clusters.

From the evaluation of the cluster with Calinski-Harabasz index. The number of clusters for a normal class is obtained in two clusters, suspect class with two clusters, and pathologic class to three clusters. So, CTG data is clustering using K-Means algorithm for each class with normal class specification clustered into 2 classes, suspect class clustered into two classes, and pathologic class clustered into 3 classes.

B. Feature Reconstruction

After getting cluster for each class, feature reconstruction applied to form new features based on the results of clustering. Untested CTG data are calculated to find similarity with the clusters using membership function in formula (5), then getting the pattern using formula (6). Pattern of each cluster is a new abstract feature from CTG with the number of features is sum of optimum number of each class. This feature reconstruction process data feature CTG from twenty one to seven features. CTG data with these seven features will be processed into training to produce a classifier.

C. SVM

Experiments performed with 10-cross validation on the CTG Data. Overall class mean accuracy obtained to 90.64% with 7 abstract features. Accuracy is calculated as follows:

$$accuracy = \frac{prediction}{actual} \times 100\% \qquad (9)$$

Where, accuracy is the percentage of correctly predicted classification results compared with the actual classification.

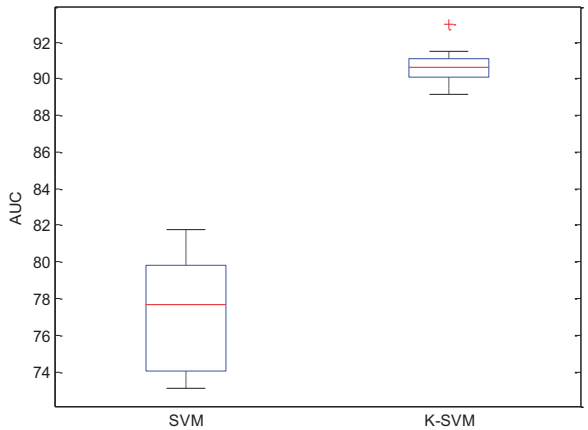

Fig. 4. AUC comparison between SVM and K-SVM

To compare the performance of K-SVM with SVM can be seen in Fig. 4. The figure shows Area Under Curves (AUC) of SVM and K-SVM based on 10 cross validation. K-SVM shows the results of greater accuracy and more stable than SVM although K-SVM feature reduces from twenty one to seven features.

Fig. 5 shows the results of K-SVM classification compared with SVM for each class. Can be seen in the figure above, there is fairly large gap between accuracy using SVM and Hybrid K-SVM. In a normal class, 1510 are classified correctly as normal classes with K-SVM method from the 1655 of data CTG fetal state which is normal, while the SVM correctly classified in 1296 as a normal class. In suspect class, K-SVM was able to classify 267 of the data as a suspect class where the baseline SVM only able to predict 228 of 295 suspect class. Similarly, in the pathologic class SVM classifies the data 107 of 176 pathologic data, while K-SVM classification results 43 of data better than the SVM. Overall, the Hybrid K-SVM has an average accuracy of 90.64% with 10 cross validation, while SVM has an average accuracy of 76.72%. From these results, a hybrid K-SVM has a better accuracy performance compared to SVM.

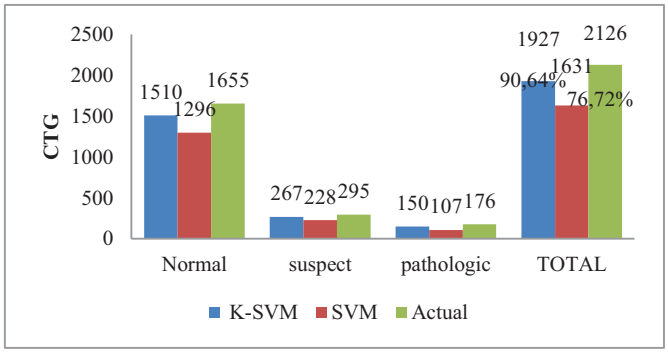

Fig. 5. Hybrid K-SVM Classification Result

D. Summary

Based on the final results, hybrid K-SVM algorithm is able to discover hidden patterns from each class in the CTG data that is class of normal fetal state, suspect, and pathologic, and reconstructs feature CTG of 21 features into the membership function of the pattern that has been previously found reduces to 7 features.

Feature extraction is not a major factor for classifying the data; however, feature extraction is very important in the classification, which in this study demonstrated the feature extraction can improve the accuracy of more than 15%.

With Hybrid K-SVM performed feature extraction process and reconstruction, this algorithm able to increase the classification results because the data with less informative are eliminated to reduce the dimensions of features of the input data and reconstruct it into a new feature to support the learning and produce optimal classifier, which is able to classify with high accuracy.

 978-1-5090-0363-1/15/$31.00

IV. Conclusions and Future Works

In this study, data mining using Hybrid K-SVM is conducted to classify Cardiotocography (CTG) data into normal, suspect, and pathologic class. This method extracting feature patterns of each class and reconstruct the data based on patterns that have been extracted by using the membership function

Feature extraction is done with K-Means clustering algorithm for finding hidden patterns in each class CTG fetal state, where each pattern representing a pattern in the class. Data reconstructed by calculating the similarity of each pattern, the similarity calculation using the membership function to determine the membership of the data to each pattern has been found in the feature extraction phase. The results of feature reconstruction produce abstract features that will be used in training to obtain classification.

In this study, feature extraction and reconstruction as a preprocessing phase in machine learning proved to be very effective to improve the classification results when compared with machine learning without preprocessing. Thus, this method is expected to help medical personnel to take medical decisions by providing interpretation of CTG readings more accurate.

Day by day, medical data will increase in number and sample data, and with the development of technology will be the greater dimension of the feature obtained. Therefore, data mining methods will always need to be explored more in depth to cope with the various problems that arise.

K-SVM reduce the dimension of features, but do not reduce the data with the same sample, it is possible to carry out further research on the sample data filtering. This will be very effective on the data with a number of very large and there are many redundant samples, thereby reducing the computational time for training

Missing value problem is another issue that needs to be resolved. In this study, researchers used data that no missing value found. However, medical data will not be separated from the missing value problem so that this issue is a challenge for researchers in the field of biomedical informatics.

References

[1] Steer P J, Has electronic fetal heart rate monitoring made a difference Semin Fetal Neonatal Med 13 (1), 2-7.

[2] Shahad Nidhal, M. A. Mohd. Ali1 and Hind Najah, “A novel cardiotocography fetal heart rate baseline estimation algorithm”, Scientific Research and Essays Vol. 5(24), pp. 4002-4010, 18 December, 2010

[3] Hakan, Sahin, and Abdulhamit Subasi. "Classification of Fetal state from the Cardiotocogram Recordings using ANN and Simple Logistic." (2012).

[4] Sundar, C., M. Chitradevi, and G. Geetharamani. "Classification of Cardiotocogram Data using Neural Network based Machine Learning Technique." International Journal of Computer Applications 47 (2012).

[5] Sundar, C., M. Chitradevi, and G. Geetharamani. "An Overview of Research Challenges for Classification of Cardiotocogram Data." Journal of Computer Science 9.2 (2013).

[6] Chudacek, V., et al. "Evaluation of feature subsets for classification of cardiotocographic recordings." Computers in Cardiology, 2008. IEEE, 2008.

[7] Prasad, Y., Biswas, K., & Jain, C. (2010). Svm classifier based feature selection using GA, ACO and PSO for sirna design. In Proceedings of the first international conference on advances in swarm intelligence (pp. 307–314).

[8] Zheng, Bichen, Sang Won Yoon, and Sarah S. Lam. "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms." *Expert Systems with Applications* 41.4 (2014): 1476-1482.

[9] Caliński, Tadeusz, and Jerzy Harabasz. "A dendrite method for cluster analysis." Communications in Statistics-theory and Methods 3.1 (1974): 1-27 (2008).

[10] Sundar, C., M. Chitradevi, and G. Geetharamani. " An Analysis on The Performance of K-Means Clustering Algorithm for Cardiotocogram Data Clustering." International Journal on Computational Sciences & Applications (IJCSA) Vo2, No.5, October 2012

[11] Wahlin, Amer I, Hellsten C, Noren H, Hagberg H, Herbst A, Kjellmer I, Lilja H, Lindoff C, Mansson M, Martensson L, Olofsson P, Sundstrom A, Marsal K. Cardiotocography only versus cardiotocography plus st analysis of fetal electrocardiogram for intrapartum fetal monitoring: a swedish randomised controlled trial. Lancet Aug 2001; 358(9281):534–538.

 978-1-5090-0363-1/15/$31.00