www.ijecs.in
*International Journal Of Engineering And Computer Science ISSN:2319-7242*
*Volume 4 Issue 7 July 2015, Page No. 13512-13517*

# Breast Cancer Using KMSVM Techniques In Wisconsin Detect Prognostic Breast Cancer Data Sets

***D. Rajakumari, C. Jayanthi***

Asst. Professor & Ph.D Research Scholar, Nandha Arts And Science College,Erode – 52, Tamil Nadu, India

Asst. Professor, Nandha Arts And Science College,

Erode – 52, Tamil Nadu, India

## ABSTRACT

Event extraction is a particularly challenging type of information extraction. Most current event extraction systems rely on local information at the phrase or sentence level. However, this local context may be insufficient to resolve ambiguity in identifying particular types of events; information from a wider scope can serve to resolve some of this ambiguity. In this paper, we first investigate how to extract supervised and unsupervised features to improve a supervised baseline system. Then, we present two additional tasks to show the benefit of wider scope features in semi-supervised learning and active learning. Experiments show that using features from wider scope can not only aid a supervised local event extraction baseline system, but also help the semi-supervised or active learning approach. The resulting efficient nugget pool is used to guide users' exploration. Among the five stages of NMS framework, we pay our main attention on solving the technical challenges existed in nugget combination and refinement.

A critical issue that makes nugget combination difficult is the distance metrics between nugget (how can we know whether two nuggets are similar or not). For chunk refinement, trying to understand what a user is looking for when a nugget was generated is a difficult job which requires effective "match" Cancer Attributes. In this thesis, we present KMSVM (K-Means Support Vector Machine) Classification solutions to both of these two challenges, and we have conducted user study to carefully compare the performances of different distance metrics between nuggets.

It is important to note that the training phase was done on 20% of the dataset, whereas the testing phase was done on the remaining 80% of the data set which are considered as unknown cases for the ALCs. The study proved that the best results obtained when the KMSVM select minimum reasonable number of features, while in the training phase the diagnostic accuracy is 0.99 and the prognostic accuracy is 0.9, and the memories ALCs achieved in the testing phase a diagnostic accuracy 0.99 and analytical accuracy 0.93.

## 1. INTRODUCTION

## BREAST CANCER MINING:

Breast cancer ranks second as a cause of cancer death in women, following closely behind lung cancer. Statistics suggest [2] the possibility of diagnosing nearly 2.5 lake new cases in India by the year 2015. Prognosis thus takes up a significant role in predicting the course of the disease even in women who have not succumbed to the disease but are at a greater risk to. Classification of the nature of the disease based on the predictor features will enable oncologists to predict the possibility of occurrence of breast cancer for a new case. The dismal state of affairs where more people are conceding to the sway of breast cancer, in spite of remarkable advancement in clinical science and therapy is certainly perturbing. This has been the motivation for research on classification, to accurately predict the nature of breast cancer[4].

Our research work mainly focuses on building an efficient classifier for the Wisconsin Prognostic Breast Cancer (WPBC) data set from the UCI machine learning repository. We achieve this by executing twenty classification algorithms via, Binary Logistic Regression (BLR), Quinlan's C4.5 decision tree algorithm (C4.5) ,Partial Least Squares for Classification (C-PLS), Classification Tree(C-RT), Cost-Sensitive Classification Tree(CS-CRT), Cost-sensitive Decision Tree algorithm(CS-MC4), SVM for classification(C-SVC), Iterative Dichomotiser (ID3), K-Nearest Neighbor(K-NN), Linear Discriminant Analysis (LDA), Logistic Regression, Multilayer Perceptron(MP), Multinomial Logistic Regression(MLR), Naïve Bayes Continuous(NBC), Partial Least Squares - Discriminant/Linear Discriminant Analysis(PLS-DA/LDA), Prototype-Nearest Neighbor(P-NN), Radial Basis Function (RBF), Random Tree (Rnd Tree), Support Vector Machine(SVM) classification algorithms. Investigate the effect of feature selection using Fisher Filtering (FF), Relief, Runs Filtering, Forward Logistic Regression (FLR), Backward Logistic Regression (BaLR) and Stepwise Discriminant (Step Disc) Analysis algorithms to enhance the classification accuracy and reduce the feature subset size. The following table1 reviews the dataset attributes and current state of research in related field of data mining [4].

TABLE I
WPBC DATASET DESCRIPTION

| *Attribute* | *Significance* | *Attribute ID* |
|---|---|---|
| ID | Unique Identity of the patient | 1 |
| Outcome | Nature of the case (R-Recurrent/N-Non- recurrent) | 2 |
| Time | TTR(Time to recur)/DFS(Disease-free Survival) | 3 |
| Radius1,2,3 | Mean of distances from centre to points on the perimeter | 4,14,24 |
| Texture1,2,3 | Standard deviation of gray-scale values | 5,15,25 |
| Perimeter1,2,3 | Perimeter of the cell nucleus | 6,16,26 |
| Area1,2,3 | Area of the cell nucleus | 7,17,27 |
| Smoothness1,2,3 | Local variation in radius lengths | 8,18,28 |
| Compactness1,2,3 | Perimeter^2 / area - 1.0 | 9,19,29 |
| Concavity1,2,3 | Severity of concave portions of the contour | 10,20,30 |
| Concave points1,2,3 | Number of concave portions of the contour | 11,21,31 |
| Symmetry1,2,3 | Symmetry of the cell nuclei | 12,22,32 |
| Fractal Dimension1,2,3 | Coastline approximation – 1 | 13,23,33 |
| Tumour | Size of the tumour | 34 |
| Lymph node | Status of the lymph node | 35 |

It presents a pseudo co-testing method for event extraction, which depends on one view from the original problem of event extraction, and another view from a coarser granularity task. Moreover, multiple selection criteria were applied to seek training examples that are informative, representative, and varied.

## 2. NUGGETS MANAGEMENT SYSTEM

In this work, we design, implement and evaluate a novel analysis-guided exploration system, called the Nuggets Management System (NMS), which leverages the collaborative effort of human intuition and computational analysis to facilitate the process of visual analytics. Specifically, NMS first extracts nuggets based on both the explicit and implicit indication of users' interest. To eliminate possible redundancy among the collected nuggets, NMS combines similar nuggets by conducting nugget clustering. Then, data mining techniques are applied to refine the nuggets and thus improve their accuracy in capturing patterns present in the datasets. We also provide a rich set of functionalities to manage the nuggets. With them, nuggets can be maintained automatically.

• We introduce a novel framework of analysis-guided visual exploration, which facilitates visual analytics of multivariate data.

We present a nugget combination solution that effectively reduces the potential redundancy among nuggets. We design a novel distance metric which effectively capture the distances between nuggets, and our user study shows that it matches well with users' intuition.

• We present a nugget refinement solution, which utilizes data analysis techniques to improve accuracy of the nuggets in capturing patterns in datasets. This is a novel approach that leverages the advantages of both human intuition and computational analysis. It not only improves the accuracy of user's discoveries, but also avoids expensive global data mining process.

• We develop tools for the management and support of visual exploration based on a learned nugget pool.

• We apply the above techniques of NMS to J free Chart and Weka tool, a freeware multivariate data visualization tool.

• We describe user studies evaluating the effectiveness of NMS. The user study demonstrates that NMS is able to

enhance both the efficiency and accuracy of knowledge discovery tasks.

## 3. KMSVM AND ITS USES:

The objective of Medical intelligence (MI) is to make well-informed Medical decisions by building both succinct and accurate models based on massive amounts of practical data. There are many kinds of models built for different practical problems, such as classifiers and repressors. This paper mainly discuses the related issues about the design of classifiers applied into BI systems.

SVM can build classifiers with high testing accuracy; the response time of SVM classifiers still needs to improve when applied into real-time MI systems. Two elements affecting the response time of SVM classifiers are the number of input variables and that of the support vectors. While Vilene et al. (2001) improve response time by selecting parts of input variables; this paper tries to improve the response time of SVM classifiers by reducing support vectors. Based on the above motivation, this paper proposes a new algorithm called K-means SVM (KMSVM). The KMSVM algorithm reduces support vectors by combining the K-means clustering technique and SVM. Since the K-means clustering technique can almost preserve the underlying structure and distribution of the original data, the testing accuracy of KMSVM classifiers can be under control to some degree even though reducing support vectors could incur a degradation of testing accuracy. In the KMSVM algorithm, the number of clusters is added into the training process as the input parameter except the kernel parameters and the penalty factor in SVM. In unsupervised learning, e.g., clustering, usually the number of clusters is subjectively determined by users with domain knowledge. However, when the K-means clustering technique is combined with SVM to solve the problems in supervised learning, e.g., classification, some objective criteria independent of applications can be adopted to determine these input parameters. In supervised learning, determining the input parameters is called model selection. Some methods about model selection have been proposed, e.g., the hold-out procedure, cross-validation (Langford, 2000), and leave-one-out (Vapnik, 1998). This paper adopts the hold-out procedure to determine the input parameters for its good statistical properties and low training costs.

## 4.MODEL SELECTION USING KM-SVM:

Model selection in the KMSVM algorithm is to decide three input parameters: the RBF kernel parameter $\gamma$, the penalty factor C, and the compression rate CR in equation (4) (searching CR is equivalent to doing the number of clusters when the number of the original data is fixed). This section discuses model selection from two perspectives: the generalization accuracy and response time of classifiers applied into real-time MI systems. Tradeoff of generalization accuracy and response time determines the values of input parameters. CR = No. of original data/No. of clusters (4) In model selection, generalization accuracy is usually estimated by some procedures, e.g., hold-out, k-fold cross-validation, and the leave-one-out procedure, since the distribution of the original data is often unknown and the actual error cannot be calculated. The hold-out procedure divides the data into two parts: the training set on which classifiers are trained, and the testing set on which the testing accuracy of classifiers is measured (Langford, 2000). The k-fold cross-validation procedure divides the data into k equally sized folds. It then produces a classifier by training on k – 1 folds and testing on the remaining fold. This is repeated for each fold, and the observed errors are averaged to form the k-fold estimate (Langford, 2000). This procedure is also called leave-one-out when k is equal to the number of trained data. This paper recommends and adopts the hold-out procedure to determine the input parameters in the KMSVM algorithm (and SVM) for two reasons. Regardless of the learning algorithms, Hoffding bounds can guarantee that with high probability discrepancy between estimated error (testing error) and true error (generalization error) will be small in the hold-out procedure. Moreover, it is very time consuming for the k-fold cross-validation or the leave-one-out procedure to estimate generalization accuracy in training large-scale data. Hence, the hold-out procedure is a better choice from the perspective of training costs[6]. The response time of KMSVM (or SVM) classifiers is affected by the number of support vectors according to the representation of KMSVM (SVM) classifiers in equation. Hence, model selection is implemented according to the

tradeoff between testing accuracy and response time (the number of support vectors).

## 5. PROBLEM FORMULATION

Generally, the actual refinement is divided into two phases, called the match and the refine phases. In NMS, such a natural "evolution" process of nuggets can also be controlled by users. NMS allows users to cease, quicken, or slow down the "evolution" by setting different parameters, such as initial "vitality", fading rate, and increasing rate. Besides such macro control, users can also directly manipulate any individual nugget. For example, users can mark a nugget as crucial, indicating that it should never be expired from the system. They could also directly delete some useless nuggets.

Match phase: In this phase, we aim to match the identified nuggets with patterns "around them" within the data space. In other words, our goal is to determine which patterns users were searching for when these specific nuggets were made. Briefly, the concept of "Match" is used to judge whether some data patterns or the major parts of these patterns primarily contribute to a nugget. If it is the case, we call the nugget and these patterns "matched". It shows a good example of a "match" between a nugget and a cluster pattern in the dataset. The specific techniques utilized to calculated how much a nugget is "matched" with the patterns around it will be described in our project [7].

Refinement Phase: The match phase reveals to us what type of patterns that a user was searching for. With this knowledge, we can finish nugget refinement using the two steps of splitting (if necessary) and modification. These two steps will make each nugget a perfect representative of a single pattern.

## 6. PROPOSED METHODOLOGY

### 6.1 FEATURE REDUCTION BY KMSVM

Feature reduction applies a mapping of the multidimensional space into a space of lower dimensions. Feature extraction includes features construction, space dimensionality reduction, sparse representations, and feature selection all these techniques are commonly used as preprocessing to machine learning and statistics tasks of prediction, including pattern recognition. Although such problems have been tackled by researchers for many years, there has been recently a renewed interest in feature extraction. The feature space having reduced features truly Contributes to classification that cuts preprocessing costs and minimizes the effects of the 'peaking phenomenon' in classification. Thereby improving the overall performance of classifier based intrusion detection systems. The commonly used dimensionality reduction methods include supervised approaches such as linear discriminant analysis (LDA), unsupervised ones such as and additional spectral and manifold learning methods.

KMSVM is a linear transformation with linear ortho normal basis vectors; it can be expressed by a translation and rotation. It convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. If we consider the two dimensional case the basic principle of this transformation.
Classification Using KMSVM is the classification table algorithm used by the immune system to describe the basic features of an immune response to an antigenic stimulus. Clonal selection establishes the idea that only cells that recognize the antigens will proliferate where the rest will not. The most triggered cells selected as memory cells for future pathogens attacks and the rest mature into antibody secreting cells called plasma cells. Clonal selection in AIS is the selection of a set of artificial lymphocytes ALCs with the highest calculated a□nifty with a non-self pattern. The selected ALCs are then cloned and mutated in an attempt to have a higher binding a□nifty with the presented no self pattern. The mutated clones compete with the existing set of ALCs, based on the calculated between the mutated clones and the non-self pattern, for survival to be exposed to the next non-self pattern.

The selection of a lymphocyte by a detected antigen for clonal proliferation inspired the modeling of KMSVM. The algorithm below summarizes the general KMSVM for pattern recognition tasks. When applied to pattern matching, a set of antigens, G, to be matched. The task of KMSVM is to then produce a set of memory ALC M that matches the members in G.

## 6 .2 STEPS FOR KM-SVM

**Step 1:** three input parameters are selected: the kernel parameter γ, the penalty factor C, and the
Compression rate CR

**Step 2:** the K-means clustering algorithm is run on the original data and all cluster centres are regarded as the compressed data for building classifiers

**Step 3:** SVM classifiers are built on the compressed data

**Step 4:** three input parameters are adjusted by the Cancer Attribute searching strategy proposed in this paper according to a tradeoff between the testing accuracy and the response time

**Step 5:** return to Step 1 to test the new combination of input parameters and stop if the combination is acceptable according to testing accuracy and response time Step 6: KMSVM classifiers are represented as the formula in equation (2).

## 6.3 SYSTEM IMPLEMENTATION

This project implemented for Breast Cancer diagnostic and prognostic results for an immerge between immune-computing and features reduction. Where an immune-computing is one of the newest directions in bio-inspired machine learning and has very fruitful successes in different area.

The classification selection theory is one of the first applied theories in KMSVM, and in this paper supported with features reduction technique KMSVM as a first step before the start of immune defense.

The presented results are very good but the false alarm it must be improved using optimization algorithms. As future work it must be make more importance to the parameters values and propose a new method to search the best values of these ones in order to across the performance of this hybrid collection of KMSVM and features reduction techniques.

## 7. CONCLUSION

We can conclude that information from wider scope can aid event extraction based on local features, including different learning methods: supervised, semi-supervised, or active learning. Also, there are different ways to extract wider scope information from different levels, which need to be further explored. For example, can the different features be combined together, and which combination is the best to find disease diagnosis? Can wider scope features help other NLP Natural Language Processing tasks, like relation extraction, named entity extraction, etc.

## 8. REFERENCES


[1] A. Koufakou and M. Georgiopoulos, "A Fast Outlier Detection Strategy for Distributed High-Dimensional Data Sets with Mixed Attributes," Data Mining and Knowledge Discovery, vol. 20, no. 2, special issue SI, pp. 259-289, Mar. 2010.

[2] R.A. Weekley, R.K. Goodrich, and L.B. Cornman, "An Algorithm for Classification and Outlier Detection of Time-Series Data,"

J. Atmospheric and Oceanic Technology, vol. 27, no. 1, pp. 94-107, Jan. 2010.

[3] M. Ye, X. Li, and M.E. Orlowska, "Projected Outlier Detection in High-Dimensional Mixed-Attributes Data Set," Expert Systems with Applications, vol. 36, no. 3, pp. 7104-7113, Apr. 2009.

[4] K. McGarry, "A Survey of Interestingness Measures for Knowledge

Discovery," Knowledge Eng. Rev., vol. 20, no. 1, pp. 39-61, 2005.

[5] L. Geng and H.J. Hamilton, "Interestingness Measures for Data Mining: A Survey," ACM Computing Surveys, vol. 38, article http://doi.acm.org/10.1145/1132960.1132963, Sept. 2006.

[6] E. Triantaphyllou, Data Mining and Knowledge Discovery via Logic-Based Methods. Springer, 2010.

[7] E.M. Knorr, R.T. Ng, and V. Tucakov, "Distance-Based Outliers: Algorithms and Applications," VLDB J., vol. 8, no. 3/4, pp. 237-253, 2000.

[8] D. Hawkins, Identification of Outliers (Monographs on Statistics and

Applied Probability). Springer, http://www.worldcat.org/isbn/041221900X, 1980.

[9] F. Angiulli and C. Pizzuti, "Outlier Mining in Large High-Dimensional Data Sets," IEEE Trans. Knowledge Data Eng., vol. 17,no. 2, pp. 203-215, Feb. 2005.

[10] Y. Tao, X. Xiao, and S. Zhou, "Mining Distance-Based Outliers from Large Databases in Any Metric Space,"

Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),T. Eliassi-Rad, L.H. Ungar, M. Craven, and D. Gunopulos, eds.,pp. 394-403, 2006.

[1]E.M. Knorr, R.T. Ng, and V. Tucakov, "Distance-Based Outliers:Algorithms and Applications," VLDB J., vol. 8, no. 3/4, pp. 237-253, 2000.

[12] M.M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," SIGMOD Record,vol. 29, no. 2, pp. 93-104,2000.

[13] Charu C.Aggarwal and Philip S.Yu,"Outlier Dection for High Dimensional Data" Data Mining and Knowledge Discovery, vol. 11, no. 6,2001

[14] N. Panda, E.Y. Chang, and G. Wu, "Concept Boundary Detection for Speeding Up SVMs," Proc. 23rd Int'l Conf. Machine Learnin g(ICML), W.W.Cohen and A. Moore eds., vol. 148, pp. 681-688,2006.

[15] Y. Tao, X. Xiao, and S. Zhou, "Mining Distance-Based Outliers from Large Databases in Any Metric Space," Proc. 12th ACMSIGKDD Int'l Conf.Knowledge Discovery and Data Mining (KDD),T. Eliassi-Rad, L.H. Ungar, M. Craven, and D. Gunopulos, eds.,pp. 394-403, 2006.