ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 4, pp: (271-278), Month: October - December 2014, Available at: www.researchpublish.com

A Novel Approach for Identifying Critical Nuggets Using PCA with ANOVA

[1]K.Devi, [2]M.Moorthi

[1] M.C.A., Research Scholar, Department of Computer Science, Kongu Arts and Science College, Erode, Tamilnadu.
[2] M.C.A., M.Phil., P.hD., Assistant Professor, Department of Computer Application, Kongu Arts and Science College, Erode, Tamilnadu.

Abstract*: Breast cancer ranks second as a reason behind cancer death in girls, following closely behind carcinoma. Statistics counsel the chance of designation nearly two lakhs new cases in Bharat by the year 2015. Prognosis so takes up a big role in predicting the course of the illness even in girls UN agency haven't succumbed to the illness however area unit at a larger risk to. Experiments show that exploitation options from wider scope cannot solely aid a supervised native event extraction baseline system, however conjointly facilitate the semi-supervised or active learning approach. A vital issue that produces hunk combination tough is that the distance metrics between hunk (how will we all know whether or not 2 nuggets area unit similar or not). For hunk refinement, attempting to know what a user is probing for once a hunk was generated may be a tough job which needs effective "match" heuristics. during this thesis, we tend to gift PCA (Principle element Analysis) with analysis of variance Classification solutions to each of those 2 challenges, and that we have conducted user study to rigorously compare the performances of various distance metrics between nuggets. it's necessary to notice that the coaching section was done on 2 hundredth of the dataset, whereas the testing section was done on the remaining eightieth of the info set that area unit thought of as unknown cases for the ALCs. The study proved that the most effective results obtained once the PCA choose minimum cheap range of options, whereas within the coaching section the diagnostic accuracy is zero.99 and also the prognostic accuracy is zero.9, and also the reminiscences ALCs achieved within the testing section a diagnostic accuracy zero.99 and prognostic accuracy zero.93.

***Keywords*: PCA-Principle Component analysis.**

I. INTRODUCTION

A common problem in data mining is that of automatically finding outliers or anomalies in a data set. Outliers are those points that are highly unlikely to occur given a model of the data. Since outliers and anomalies are rare, they can be indicative of bad data, faulty collection, or malicious content. There are several approaches to outlier detection. One approach is that of model-based outlier detection, where the data is assumed to follow a parametric distribution. Such approaches do not work well in even moderately high dimensional spaces and finding the right model is often a difficult task in its own right. To overcome these limitations, researchers have turned to various non-parametric approaches that use a point's distance to its nearest neighbor as a measure of unusualness. We further improve the scaling behavior of distance-based outlier detection on large, high-dimensional data sets.

A. Breast Cancer Mining:

Breast cancer ranks second as a cause of cancer death in women, following closely behind lung cancer. Statistics suggest [7-8] the possibility of diagnosing nearly 2.5 lakhs 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.

The research work mainly focuses on building an efficient classifier for the Wisconsin Prognostic Breast Cancer (WPBC) data set from the UCI machine learning repository [9-12]. We achieve this by executing twenty classification algorithms viz, 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), investigate the effect of feature selection using Fisher Filtering (FF), ReliefF, Runs Filtering, Forward Logistic Regression (FLR), Backward Logistic Regression (BaLR) and Stepwise Discriminate (Step Disc) Analysis algorithms to enhance the classification accuracy and reduce the feature subset size.

B. Unsupervised Features:

The third category extracts unsupervised features from distributions in large-scale untagged corpora. Such studies include Riloff (1996), Yangarber et al. (2000). These features are used when there is not much training data, or the training and testing data has different distribution.

In this thesis, 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 (self training) and active learning (co-testing)

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

.

The table shows that, the description of he attributes which used in the dataset. Where attribute, significance and attribute ID. Compared to a supervised multi-label classifier, the unsupervised approach can achieve comparable, even better, results. Also, we do not limit our study to the corpus from the standard ACE evaluation, which is preselected, but also investigate its performance on a regular newswire corpus. To presents a self-training process for event extraction. Event-centric entity-level Information Retrieval (IR) techniques were incorporated to provide topic-related document clusters.

C. Automatic Content Extraction (Ace) Evaluation:

ACE began in 2000 after MUC. "The objective of the ACE program is to develop automatic content extraction technology to support automatic processing of human language in text form from a variety of sources (such as newswire, broadcast conversation, and weblogs). ACE technology R&D is aimed at supporting various classification, filtering, and selection applications by extracting and representing language content (i.e., the meaning conveyed by the data).

a) **Entity Detection and Recognition (EDR)** is the core annotation task of ACE, providing the foundation for all remaining tasks. This ACE task identifies seven types of entities: Person, Organization, Location, Facility, Weapon, Vehicle and Geo-Political Entity (GPE). Each type is further divided into subtypes. Annotators tag all mentions of each entity within a document, whether named, nominal (common noun) or pronominal.

b) To introduce a novel framework of analysis-guided visual exploration, this facilitates visual analytics of multivariate data.

c) To 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

II. RELATED WORK

An algorithm to perform outlier detection on time-series data is developed, the intelligent outlier detection algorithm (IODA). This algorithm treats a time series as an image and segments the image into clusters of interest, such as ''nominal data'' and ''failure mode'' clusters. The algorithm uses density clustering techniques to identify sequences of coincident clusters in both the time domain and delay space, where the delay space representation of the time series consists of ordered pairs of consecutive data points taken from the time series. ''Optimal'' clusters that contain either mostly nominal or mostly failure-mode data are identified in both the time domain and delay space.

A best cluster is selected in delay space and used to construct a ''feature'' in the time domain from a subset of the optimal time-domain clusters. Segments of the time series and each datum in the time series are classified using decision trees. Depending on the classification of the time series, a final quality score (or quality index) for each data point is calculated by combining a number of individual indicators. The performance of the algorithm is demonstrated via analyses of real and simulated time-series data.

III. SYSTEM DESIGN

A. Existing System

In the existing system, this cancer affects one in eight women during their lives. It occurs in both men and women, although male breast cancer is rare. Breast cancer is a malignant tumor that has developed from cells of the breast. Although scientists know some of the risk factors (i.e. ageing, genetic risk factors, family history, menstrual periods, not having children, obesity) that increase a woman's chance of developing breast cancer, they do not yet know what causes most breast cancers or exactly how some of these risk factors cause cells to become cancerous. Research is under way to learn more and scientists are making great progress in understanding how certain changes in DNA can cause normal breast cells to become cancerous.

Some kind of patient's cancer identification is insufficient to predict it. The used data source is Wisconsin Breast Cancer Dataset (WBCD) taken from the University of California at Irvine (UCI) Machine Learning Repository. This data set is taken from fine needle aspirates from human breast tissue was commonly used among researchers who use machine learning (ML) methods for breast cancer classification. There are 699 records in this dataset. Each record in the database has nine attributes. The nine attributes detailed in Table 1 are graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 4, pp: (271-278), Month: October - December 2014, Available at: www.researchpublish.com

TABLE I
Wisconsin breast cancer data description of attributes

Wisconsin breast cancer data description of attributes				
Attribute number	Attribute description	Values of attributes	Mean	Standard deviation
1	Clump Thickness	1–10	4.42	2.82
2	Uniformity Of Cell Size	1–10	3.13	3.05
3	Uniformity Of Cell Shape	1–10	3.2	2.97
4	Marginal Adhesion	1–10	2.8	2.86
5	Single Epithelial Cell Size	1–10	3.21	2.21
6	Bare Nuclei	1–10	3.46	3.64
7	Bland Chromatin	1–10	3.43	2.44
8	Normal Nucleoli	1–10	2.87	3.05
9	Mitoses	1–10	1.59	1.71
N = 699 Observations, 241 Malignant and 458 Benign.				

B. Proposed System:

The proposed system design is diagrammatically presented in Fig 1. The data mining framework for the classifier is viewed from the perspective of both the training/learning phase and the test phase. The dataset is visualized and pre-processed before applying any of the data mining techniques. The training phase makes the learning process complete by generating all possible rules for classification after performing feature relevance followed by classification. The test phase determines the accuracy of the classifier when presented with a test data (unseen breast cancer case) and by viewing the returned class label. This work use PCA as an aided tool for immune cells in the selection for the most important features that can detect the cancer and forward them for the immune system in training phase which generates artificial lymphocytes ALCs and save them as immune memory. 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 PCA 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 prognostic accuracy 0.92.

IV. MODULES

A. Data Visualization and Pre-processing

The Wisconsin Prognostic Breast Cancer dataset is downloaded from the UCI Machine Learning Repository website [9] and saved as a text file. This file is then imported into Excel spreadsheet and the values are saved with the corresponding attributes as column headers. The missing values are replaced with appropriate values. The ID of the patient cases does not contribute to the classifier performance. Hence it is removed and the outcome attribute defines the target or dependant variable thus reducing the feature set size to 33 attributes. The algorithmic techniques applied for feature relevance analysis and classification are elaborately presented in the following sections.

a) Feature Selection Algorithms

The generic problem of supervised feature selection [19] can be outlined as follows. Given a data set {(xi, yi)} ni=1 where xi ∈ Rd and yi ∈ {1, 2…c}, aim to find a feature subset of size m which contains the most informative features. The two well-performing feature selection algorithms on the WPBC dataset are briefly outlined below.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 4, pp: (271-278), Month: October - December 2014, Available at: www.researchpublish.com

b) Fisher Filtering

It is termed Univariate Fisher's ANOVA ranking [20]. It is a supervised feature selection algorithm that processes the selection independently from the learning algorithm. It follows a filtering approach that ranks the input attributes according to their relevance. A cutting rule enables the selection of a subset of these attributes. It is required to define the target attribute which in this domain of research applies to the nature of the breast cancer (recurrent/non- recurrent) and the predictor attributes. After computing the Fisher score [21-22] for each feature, it selects the top-m ranked features with large scores. The next subsection directs focus on another technique of feature selection based on logistic regression.

B. Leverage Backward Logistic Regression Nuggets analysis

When the number of descriptors is very large for a given problem domain, a learning algorithm is faced with the problem of selecting a relevant subset of features Backward regression includes regression models in which the choice of predictor variables is carried out by an automatic procedure. The iterations of the algorithm for logistic regression are given in steps as stated as follows.

Step 1: The feature set with all 'ALL' predictors.

Step 2: Eliminate predictors one by one.

Step 3: 'ALL' models are learnt containing 'ALL-1' descriptor each.

These iterations are further continued till either a pre-specified target size is reached or the desired performance statistics (classification accuracy) is obtained. After feature relevance, we classify the nature of the breast cancer cases in the Wisconsin Prognostic Breast Cancer dataset using twenty classification algorithms. The best performing algorithms are described in the following section.

C. Feature Reduction by PCA

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 pre-processing costs and minimizes the effects of the 'peaking phenomenon' in classification. Thereby improving the overall performance of classifier based intrusion detection systems.

V. ALGORITHM

A. Finding Principle Component Analysis Main Attributes

An observation is excluded only in the calculation of covariance or correlation between two variables if missing values exist in either of the two variables for the observation.

Cancer and datavectors are calculated from the covariance or correlation matrix *S*.

$$SP = PD$$

where *P* is a *p* by *p* matrix and *D* is a diagonal matrix with diagonal elements $\lambda_i \; i = 1, 2, \ldots, p$.

- **Cancer**

λ_i is the *i*th datavalue for the *i*th principal component. And cancers are sorted in descending order.

Note that cancer can be negative for missing values excluded in a pairwise way, which will make no sense for principal components. Origin sets the loading and scores to zeros for a negative datavalue.

- **Datavectors**

Each column in *P* is the datavector corresponding to the datavalue or principal component.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 4, pp: (271-278), Month: October - December 2014, Available at: www.researchpublish.com

Note that the datavector's sign is not unique; Origin normalizes its sign by forcing the sum of each column to be positive.

- **Scores**

$$V = X_0 P$$

where X_0 is the matrix X with each column's mean subtracted from each variable.

Scores will be missing values corresponding to an observation containing missing values.

Note that variance of scores for each principal component may not equal its corresponding datavalue for this method.

- **Standardized Scores**

Scores for each principal component are scaled by the square root of its datavalue.

ANOVA tests the equality of the remaining *p-k* cancer. It is available only when analysis $H_0 : \lambda_{k+1} = \lambda_{k+2} = ... = \lambda_p k = 0, 1, ..., p-2$ matrix is covariance matrix.

It approximates a $X2$ distribution with $\frac{1}{2}(p-k-1)(p-k+2)$ degrees of freedom.

$$(n-1-(2p+5)/6)\left\{ -\sum_{i=k+1}^{p} \log(\lambda_i) + (p-k)\log\left(\sum_{i=k+1}^{p} \lambda_i/(p-k)\right)\right\}$$

VI. SYSTEM ARCHITECTURE

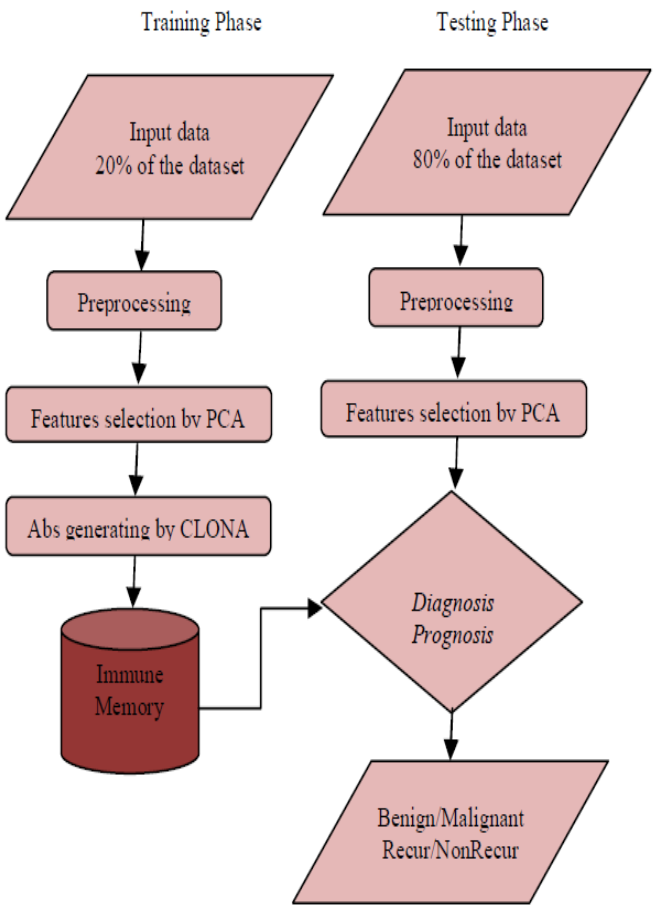

Figure 1

VII. CONCLUSION

In this project we have considered the Wisconsin Prognostic Breast Cancer (WPBC) dataset for creating an efficient Classifier since it is highly essential in any clinical investigation to determine the nature of a disease, especially a life threatening ailment like cancer. The results of classification after feature selection are clearly outlined in this project with necessary results. This will make it easier for Oncologists to differentiate a good prognosis (non-recurrent) from a bad one (recurrent) and classify any new breast cancer dataset as being of a recurrent nature or non-recurrent one. Further accurate classification would enable clinicians to propose drugs for a new patient based on whether his/her features correspond to a good or bad prognosis. According to our findings, Fisher Filtering, Backward Logistic Regression, Stepwise LBLR Classification Analysis and PCA algorithms have performed well in terms of improving classifier accuracy on this dataset. PCA Tree and LBLR classification algorithms have produced 100 percent accuracy in classifying the Wisconsin Prognostic Breast Cancer dataset. We also affirm that the Quinlan's C4.5 algorithm is the best performing classification algorithm on the WPBC dataset in terms of storage and classification accuracy since the decision tree generated is smaller and it also provides 100 percent classification accuracy.

Future Work

From the above studies, 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 diagonsis? Can wider scope features help other NLP Natural Language Processing tasks, like relation extraction, named entity extraction, etc.?

REFERENCES

[1] Ali A., Shamsuddin S. M., Ralescu A. L., and Visa S., "Fuzzy Classifier for Classification of Medical Data", 2011 11th International Conference on Hybrid Intelligent Systems (HIS), 2011 IEEE, pp. 173-178.

[2] Anagnostopoulos I., Anagnostopoulos C., Rouskas A , Kormentzas G., and Vergados D., "The Wisconsin Breast Cancer Problem: Diagnosis and DFS time prognosis using probabilistic and generalised regression neural classifiers", DRAFT VERSION of paper to appear at the Oncology Reports, special issue Computational Analysis and Decision Support Systems in Oncology, last quarter 2005, IVSL.

[3] Balakrishnan S., Narayanaswamy R., and Paramasivam I., "An Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets", International Journal of Computer Applications (0975 – 8887) Volume 29– No.5, September 2011, pp. 1-6, IVSL.

[4] Engelbrecht A. P., 2007, "Computational Intelligence An Introduction", Second Edition.

[5] Dasgupta D. and Niño L. F., 2009, "Immunological Computation Theory and Applications", Auerbach Publications.

[6] Gupta S., Kumar D., and Sharma A., "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2 No. 2 Apr-May 2011, pp.188-195, IVSL.

[7] Karabatak M., and Ince M. C., "An expert system for detection of breast cancer based on association rules and neural network", ScienceDirect, 2008 Elsevier, IVSL.

[8] Khelil H, and Benyettou A., "Artificial Immune Systems For Illnesses Diagnostic", Ubiquitous Computing and Communication Journal, Volume 3 Number 4, 2007, pp. 88-93.

[9] lakhina S., Joseph S., and Verma B., "Feature Reductionusing Principal Component Analysis for Effective Anomaly– Based Intrusion Detection on NSL-KDD", International Journal of Engineering Science and Technology, Vol. 2(6), 2010, pp.1790-1799, IVSL.

[10] Ludwig S. A. and Roos S., "Prognosis of Breast Cancer using Genetic Programming", Department of Computer Science, University of Saskatchewan, Canada,ludwig@cs.usask.ca.

[11] Nabil E., Badr A., Farag I., and Osama M., "A Hybrid Artificial Immune Genetic Algorithm with Fuzzy Rules for Breast Cancer Diagnosis", INFOS2008, March 27-29, 2008 Cairo-Egypt.

[12] RADI A., Kartit A., REGRAGUI 3B., El Marraki 4M., ABOUTAJDINE D., and RAMRAMI A.," On the Three Levels Security Policy Comparison between PCA and SVM",IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.2, February 2011, pp.69-79.

[13] Salama G. I., Abdelhalim M.B., and Zeid M. A., "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", International Journal of Computer and Information Technology (2277 – 0764), Volume 01–Issue01, September 2012, pp. 36-43.

[14] Shyu M., Chen S., Sarinnapakorn K., and Chang L.,"A Novel Anomaly Detection Scheme Based on Principal Component Classifier", in Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM'03), 2003, pp.172-179, IVSL.