STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720(online)
Volume 2, Issue 2
www.stmjournals.com

Chi-Square Statistic and Principal Component Analysis Based Compressed Feature Selection Approach for Naïve Bayesian Classifier

Biprodip Pal*, Sadia Zaman, Md. Abu Hasan, Md. Al Mehedi Hasan, Firoz Mahmud
Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Bangladesh

Abstract

Many of the machine learning algorithms are based on an assumption of attribute independency and often used in domains where the assumption doesn't hold true. Naïve Bayesian (NB) classifier makes assumption that all the features are conditionally independent given the class labels; In this paper, attribute dependencies were analyzed using Chi-Square test and the Principal Component Analysis (PCA) was carried out on the whole dataset to get a set of features. We have also applied PCA on the independent attributes of the data with a view to get a more compressed set of features that may lead to reliable accuracy for NB Classifier. The performance of the classifier was experimented for the combined approach as well as individual Chi-Square and PCA only approach with variation in dataset sizes. It was found that, for the used dataset, reduced dimensionality of the dataset according to Chi-Square independency test as well as the combined approach has come out with much better performance than PCA only approach, but considering the time, the combined approach is better.

Keywords: *Chi-square test, machine learning, naïve bayesian classifier, principal component analysis*

****Author for Correspondence*** E-mail: biprodip.cse@gmail.com

INTRODUCTION

In machine learning, classification is the process of assigning a new instance to one of a given set of classes. Decision tree classifiers, Rule-based classifiers, Bayesian classifiers, Support vector machines (SVM), artificial neural networks and Ensemble methods are most prominent classification methods. In this paper, we have worked with Naïve Bayesian classifier. This classifier is an example of supervised learning. In supervised learning, the training set consists of a set of n ordered pairs $(x_1,y_1),(x_2,y_2),\ldots, (x_n,y_n)$ where each x_i is some measurement or set of measurements of a single example data point, and y_i is the label for that data point [1]. Naïve Bayesian classifier is very popular for its simplicity and robustness. It is a statistical classifier, based on Bayes theorem. It has its own appeal because it is easy to understand and the induction of theses classifiers is extremely fast, requiring only a single pass through the data if all attributes are discrete. This classifier is very robust to irrelevant attributes and classification takes into account evidence from many attributes to make the final prediction, a property that is useful in many cases where there is no "main effect" [2, 3]. Performance of classification algorithms are often affected by the features used from the training set. Among feature selection techniques, statistical techniques involves mostly correlation analysis of attributes [4], on the other hand Conceptual techniques involves use of Decision tree, Decision rules and similar machine learning tools to get the desired feature subset [5]. Combined conceptual and statistical approaches are also used besides Rough Set, as well as evolutionary techniques like Genetic algorithm [6–8].

But Bayesian classifier assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. But most of the times, this assumption is violated; because usually the attributes are dependent [9]. The

purpose of this study was to see how much variation it brings on the Bayesian classifier's accuracy rate, when only independent attributes were used to train and test the classifier. The first task was finding out the independent attributes from the used dataset. For that, Chi-Square test of independence was used. But Chi-Square Approach (CSA) is not applicable on continuous attributes. So, the continuous attributes were discretized using a simple discretization method, named equal width discretization (EWD).

Appplying over different datasets, we have achived a compressed set of data using PCA. Then, the Bayesian classifier was tested separately on independent attributes and on whole dataset which contains both dependent and independent attributes. After that, their corresponding accuracy rates were compared.

NAÏVE BAYESIAN CLASSIFIER

A classifier is a function f apping input feature vector X to output class labels $\{C_1,C_2,\ldots,C_m\}$ where $X=\{x_1,x_2,\ldots x_n\}$,the n measurements of a tuple for n features. Once trained, the classifier can be used to determine whether the features used contain information about the class.

Bayesian classifiers work by assuming an underlying probabilistic model, the Bayes theorem [10].By Bayes' theorem,

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (1)$$

Here,$C_1,C_2,\ldots,C_m$ are *m* numbers of classes and X is an instance. The naïve Bayesian classifier predicts that X belongs to the class C_i if and only if

$$P(C_i \mid X) > P(C_j \mid X)$$

for $1 \le j \le m, j \ne i$.The class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis.

To reduce computational cost, the naïve assumption of class conditional independence is made. It is assumed that the values of the attributes are conditionally independent of one another, given the class. Mathematically, it can be expressed as

$$P(X \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i) \quad (2)$$

Here, x_k is the value of attribute A_k for instance X. The Bayesian classifier is tolerant for noisy and incompetent data. It has been found to be comparable in performance with decision tree and selected neural network classifiers [10, 11].

EQUAL WIDTH DISCRETIZATION

Discretization is the process of converting continuous data into discrete data. Typically, a discretization process [12, 13] works as follows- first, the continuous valued attributes which are to be discretized are sorted. Then, a cut point is evaluated for splitting or adjacent intervals for merging. According to some criterion, intervals of continuous values are split or merged. Then, finally it is stopped at a point, which is called stopping criterion. The equal-width discretization (EWD) algorithm determines the minimum and maximum values of the discretized attribute and then divides the range into the user-defined number of equal width discrete intervals. This is an unsupervised discretization process. Figure 1 gives a graphical view of EWD. Suppose there are *n* training instances.To discretize an attribute X_i first the instances are sorted into ascending order. The minimum and maximum values are referred to as v_{min} and v_{max} respectively. EWD then divides the sorted attributes between v_{min} and v_{max} into *k* intervals of equal width. Thus the intervals have width, and the cut points are at $v_{min}+w$, $v_{min}+2w,\ldots, v_{min}+(k-1)w$

$$w=(v_{max}-v_{min})/k \quad (3)$$

Here, *k* is a user defined parameter [14].

CHI-SQUARE TEST

Contingency Table

When the members of a sample are doubly classified, the results may be arranged in rectangular tables. Such a table is called contingency table [15]. It is used to analyze and record the relationship between two or more categorical variables. The contents found in the rows of a contingency table are dependent upon the contents of columns.

Expected Frequency

An expected frequency is a theoretically predicted frequency obtained from an experiment. It is assumed to be true until statistical evidence indicates otherwise. In contingency table, if only the row and column

totals were known, and if it was assumed that the variables under comparison were independent, then the expected frequencies would be predicted in each cell of the table.

Observed Frequency

In contingency table, observed frequency is the actual frequency that is obtained in each cell of the table. The events being predicted must be mutually exclusive and the frequencies are from random sample [16].The differences between observed and expected frequencies suggest that the model expressed by the expected frequencies does not describe the data well.

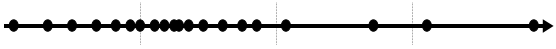

Fig. 1: *Equal width discretization [13].*

Hypothesis Test

A statistical hypothesis is an assumption about a population parameter, which may or may not be true. Hypothesis testing is the process of choosing between competing hypotheses about a probability distribution. It is a core topic in mathematical statistics. While considering a problem, the question of interest is simplified into two competing hypotheses; the null hypothesis, denoted *H0* and the alternative hypothesis, denoted *H1*. In hypothesis testing, if null hypothesis is found to be true then the alternative hypothesis is rejected and vice-versa.

Significance Level

Significance level is the probability of rejecting the null hypothesis when it is true. It is the probability of a type I error and is usually made as small as possible in order to protect the null hypothesis [16].The significance level is denoted by α

Significance Level, α= P (type I error)

Usually, α is chosen to be 0.05.

P- Value

In a statistical hypothesis test, if the null hypothesis is true, then the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone is called the probability value (p-value). It is equal to the significance level of the test for which the null hypothesis will be rejected only. When the p-value is compared to the actual significance level of a test and, if it is found smaller, then the result is significant. For example, if the null hypothesis were to be rejected at the 5% significance level, then this would be written as "$p<0$". The smaller the p-value is, the more likely the null hypothesis is going to be rejected. It indicates the strength of evidence. In Figure 2, a graphical view of p-value is given.

Degrees of Freedom

It can be found by the degrees of freedom (DF) that how many numbers in the grid are actually independent. For a Chi-Square grid, the DF is the number of cells one needs to fill in before, given the totals in the margins, one can fill in the rest of the grid using a formula. The DF for a Chi-square grid can be given as *(R-1)*(C-1)*. Here, *R* stands for rows and *C* for columns. DF in a chi-square test factors into calculations of the probability of independence. Once a Chi-square value is calculated, this value and the DF are used to decide the probability, or p-value, of independence.

Chi-Square Distribution Table

In a chi-square distribution table, the probability values are represented by the first row and the degrees of freedom by the first column. To find probability, for a given DF, one has to read across the row bellow, until the next smallest number is found. Then one has to move to the top and find the probability. Figure 3 shows the graphical view of chi-square distribution. Here, the shaded area is equal to α for $\chi^2 = \chi\alpha^2$. A portion of chi-square distribution table [17] is given in Table 1.

Independence Test

A chi-square test of independence is used to determine if two variables are related to each other using contingency table. The entries in the cells in a contingency table may be frequencies or frequencies may be transformed into proportions or percentages. However, it must be noted that, these entries are not presented in continuous measurements. Chi-squared test can be performed only on discrete data. The process of performing Chi-Square test of independence is as follows [15, 18] If there are two variables and with levels *r*

and *c* respectively, the first step of performing independence test on them is stating the hypotheses as follows.
H0: Variable *A* and variable *B* are independent.
H1: Variable *A* and variable *B* are not independent.

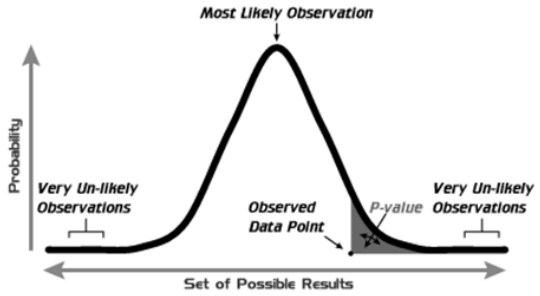

Fig. 2: *P-value (the Shaded Area).*

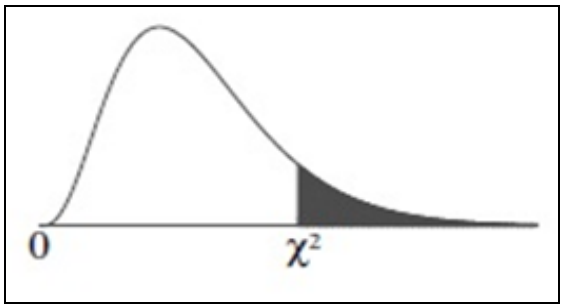

Fig. 3: *Chi-Square Distribution.*

Table 1: *Chi-Square Distribution Table.*

DF	0.050	0.025	0.010	0.001
1	3.841	5.024	6.635	10.828
2	5.991	7.378	9.210	13.816
3	7.815	9.348	11.345	16.266
4	9.488	11.143	13.277	18.467
5	11.071	12.833	15.086	20.515
…	…	…	…	…

The second step is formulating an analysis plan. The analysis plan describes how sample data can be used to accept or reject the null hypothesis. The plan should specify the significance level and test method. Any value between 0 and 1 can be used as significance level. Usually, 0.01, 0.05, or 0.10 are chosen as significance level. After that, sample data are analyzed. Using sample data, the DF, expected frequencies, test statistic, and the P-value associated with the test statistic can be found. The DF is calculated as:

$$DF = (r-1)*(c-1) \quad (4)$$

Expected frequencies can be computed, according to the following formula.

$$E_{r,c} = \frac{(n_r * n_c)}{n} \quad (5)$$

Here, $E_{r,c}$ is the expected frequency count, n_r stands for the total number of sample observations at level *r* of variable *A*, n_c for the total number of sample observations at level *c* of variable *B*. *n* is the total sample size. The value of χ^2 can be measured by the following equation:

$$\chi^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} \quad (6)$$

Where, *Or, c* is the observed frequency and *Er, c* is the expected frequency. Then, using the value of χ^2 and chi-square distribution table, p-value is calculated. To interpret the result, the p-value is compared to the significance level, and the null hypothesis is rejected when the p-value is found to be less than the significance level. Otherwise, the alternative hypothesis is rejected.

PRINCIPAL COMPONENT ANALYSIS

Principle Component Analysis (PCA) is a mathematical procedure, used in high dimension data analysis, signal processing, etc. It is a simple, non-parametric method of extracting relevant information from high-dimensional data using the dependencies between the variables to represent it in a lower-dimensional compressed form, without losing too much information.

First, it analyzes a data table of inter-correlated observations described by several dependent variables. Then it extracts the important information from the data table and expresses this information as a set of new orthogonal variables, which are called principal components.

PCA, is one of the simplest and most robust ways of doing such dimensionality reduction. It also displays the pattern of similarity of the observations and the variables by points in maps [19, 20]. Reducing the dimensionality of a data set consisting of a large number of interrelated variables, while retaining the variation present as much as possible, is the central idea of PCA.

PCA Algorithm [21]:
Suppose x1, ***x2, … x_M*** are Nx1 vectors.
Step 1: Find the mean.

$$x' = \frac{1}{M}\sum_{i=1}^{M} x_i$$

Step 2: Subtract the mean.

$$\phi_i = x_i - x'$$

Step 3: From the matrix $A = [\varphi_1 \varphi_2 ... \varphi_M]$ (N x M matrix)
Than compute Covariance Matrix

$$C = \frac{1}{M}\sum_{i=1}^{M} \varphi_i \varphi_i = AA^T$$

Step 4: Compute the Eigenvalues of C: $\lambda_1 > \lambda_2 > ... > \lambda_N$
Step 5: Compute the eigenvectors of C: $u_1, u_2, \ldots, u_n$
Since C is symmetric, $u_1, u_2, \ldots, u_n$ form a basis. (i.e., any vector x or actually $(x - x')$ can be written as a linear combination of the eigenvectors):

$$(x - x') = b_1u_1 + b_2u_2 + \ldots + b_Nu_N = \sum_{i=1}^{N} b_i u_i$$

Step 6: (Dimensionality reduction step) keep only the terms corresponding to the K largest Eigen values:

$$(x - x') = \sum_{i=1}^{K} b_i u_i$$

To select K several techniques exists. If threshold is used than, for a threshold T, K can be selected as follows.

$$\frac{\sum_{i=1}^{K} \lambda i}{\sum_{i=1}^{N} \lambda} \geq T \qquad (7)$$

EXPERIMENTAL RESULT

Used dataset

For this experiment the "thyroid disease data set" was used, from UCI machine learning repository [22]. This dataset is from Garavan Institute, Sydney, Australia; this dataset contains 6 data sets. The one used in this experiment is the thyroid dataset. The problem is to determine whether a patient referred to the clinic is hypothyroid. The output from the network is expected to be one of the three possible classes, namely: (i) normal (not hypothyroid), (ii) hyper function and (iii) subnormal function. It consists of 3772 learning examples and 3428 testing examples readily usable as ANN training and test sets. Each example has 21 attributes, 15 of which are binary and 6 are continuous. In the training set of this dataset, 2.47% data is primary hypothyroid, 5.06% - compensated hypothyroid and 92.47% is normal. While in testing set 2.13% data is primary hypothyroid, 5.16% - compensated hypothyroid and 92.71% is normal. For our experiment we have created five more datasets for training & testing by selecting patterns from the original training and testing examples respectively. The dataset label and sizes are given in Table 2.

Table 2: *Total Samples in Datasets.*

Dataset Label	No. of Train Patterns	No. of Test Patterns
D1	400	400
D2	600	600
D3	800	800
D4	1000	1000
D5	2000	2000
D6	3428	3428

Data Discretization

The data set has 6 continuous attributes. To perform Chi-square test on the dataset, these attributes were needed to be discretized. For discretization, EWD was used. First, two discretization bin was used for this method. More specifically, in equation (3), *k=2* was used. Later, the bin number was increased up to 14 since values above could not provide with significant variation in outcome and the continuous values were discretized for 2 to 14 bins. So, there are a total of 13 different discretized output set for the continuous attributes. Than Chi-square test of independence was performed on the total dataset differently for each of the 13 different cases.

Checking Dependency

Following *Chi-square* approach (CA), first DF, expected and observed frequencies were calculated. After that, the χ^2 was calculated

using equation (6). Then, with P-value the result was interpreted. This procedure was repeated for each two attributes, until all the independent attributes were found. When 2 bins were used for discretizing data, the result in Table 3 was obtained. From the second row of Table 3, it is seen that, attribute 1 has some depended attribute set {3,5,6,7,9,10,11,12,13,16,20,21}. So, our attempt is to use the dependency information, using only attribute 1 instead of {3,5,6,7,9,10,11,12,13,16,20,21}. Among the independent attribute set {2, 4, 8, 14, 15, 17, 18, 19} of 1, again the independence test was performed to select another independent attribute. The process was continued until a set of independent attributes were found. Finally, up to 14 bins were used for discretizing continuous values, and independence test were carried on for each different set of data where all the patterns consisted of discrete values. Among 21 attributes, attribute set {1, 4, 8,12 ,14} were found to be independent of each other by taking union of all the outcomes from the 13 different tests. Then a separate dataset for each of the dataset from Table 2 was formed by projecting the original ones on the new dimension.

Result Analysis

The classification accuracy of Naïve Bayesian classifier was analyzed in three different scenarios, for six different dataset. In each case the classifier was trained with training patterns from the six datasets given in Table 2 and tested on the corresponding test patterns. Firstly, features were selected from the datasets using PCA and the NB classifier was trained and tested. Secondly, the classifier was trained with features selected by applying CSA on the datasets. Thirdly, the selected features using CSA was compressed using PCA to get a more compact set of features and the classifier was allowed to train and test based on these features. Figure 4 depicts the performance of the three approaches.

To select the Principle Components we followed two approaches. Firstly using equation (7) we used a threshold of 0.8 and 0.9 and examined the result. Then since we have only six Eigen values, we also assayed by using all subsets of the eigenvectors to measure the performance. Top two Eigen values were finally selected according to performance. The only PCA based approach shows an accuracy of 82.25% for dataset D1, which rose up to 92.125% for D3. For larger dataset D6 performance dropped to a level of 64.5858%. Compared to other approaches, the PCA only approach is found to be less reliable, since for larger datasets the performance decreases, where other approaches have shown better performance.

Table 3: *Independence Test on Attributes for Two Bins.*

Attribute	Independent Attributes	Dependent Attributes
1	2,4,8,14,15,17,18,19	3,5,6,7,9,10,11,12,13,16,20,21
2	15,17,18	4,8,14,19
15	17,18	
17	18	
Final Output	1,2,15,17,18	3,4,5,6,7,8,9,10,11,12,13,14,16,19,20,21

No. significant differences were found on the overall classification performance while using the CSA, compared to that of the CA. Performance of these two approaches was same for dataset D2, D3, D4. For D1, D5 and D6 the CS approach came out with better accuracy, but the discrepancy was only 0.25, 0.05 and 0.0591 respectively. Interestingly, the performance discrepancy is not increasing as the dataset size increases. Finally, the average required times to train the classifier for CA and CSA were analyzed. We experimented only for the larger datasets from Table 2. Thus D5 and D6 was taken into consideration. In a computer with 2.5GHz Intel Core i5 processor and 1600MHz 4GB DDR3 RAM the CA came out with faster performance as expected due to the compression,

Table 4: *Average Training Time.*

Dataset Size	CSA Time(s)	CA Time(s)
2000	0.0452	0.0201
3428	0.0509	0.0232

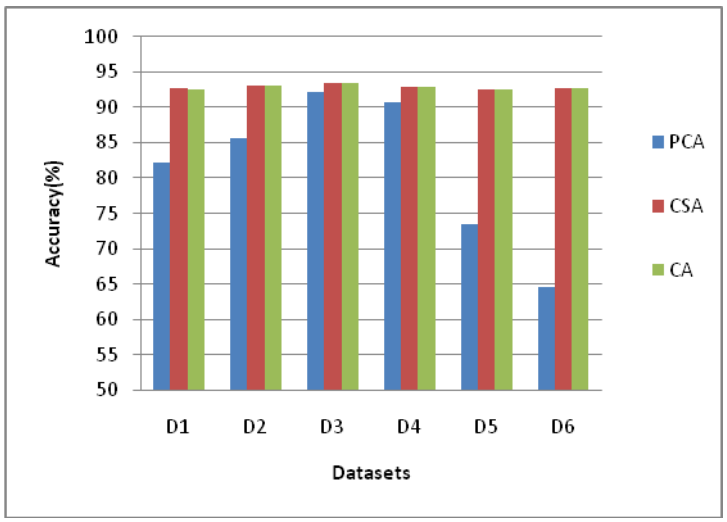

Fig. 4: *Accuracy Rate Comparison of all Approaches.*

Compared to that of CSA as shown in Table 4. Thus considering time requirement for our environment, the Chi-Square and PCA based combined approach for selecting features for NB classifier is much better compared to other two approaches.

CONCLUSION & FUTURE WORKS

The Naïve Bayesian classifier works on a simple, but comparatively intuitive concept making use of the variables contained in the data sample, by observing them individually, independent of each other. Our concern was the pre-processing phase, and the relevant effect on the classification phase. So, in future, combined approaches that may use this approach with other clustering based feature selection can also be examined for efficient selection of more robust feature subsets to use in model construction. It's capability in online adaptive systems demands evidence since this approach works surprisingly well with reduced set of features.

REFERENCES

1. Kotsiantis, S. B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica*.2007; 31: 249–268p.
2. Kohavi, R. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* Portland, OR: AAAI Press. 1996: 202–207p.
3. Qin, Biao, Yuni Xia, Fang Li. A Bayesian classifier for uncertain data. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, ACM. 2010: 1010–1014p.
4. Chan K.C.C., Wong A.K.C. A Statistical Technique for Extracting Classificatory Knowledge from Databases. *Knowledge Discovery in Databases, G. Piatetsky-Shapiro and W.J. Frawley, eds., Cambridge, Mass.*: AAAI/MIT Press. 1991: 107–123p.
5. Kaufman K., Michalski R.S., Kerschberg L. Mining for Knowledge in Data: Goals and General Description of the INLEN System. *IJCAI-89 Workshop on Knowledge Discovery in Databases, Detroit, MI*. 1989.
6. Imam I. F., Michalski R. S., Kerschberg L. Discovering attribute dependence in databases by integrating symbolic learning and statistical analysis techniques. *In Proceeding of the AAAI-93 workshop on knowledge discovery in databases*.1993.
7. Stein G, Chen B, Wu AS, Hua KA. Decision tree classifier for network intrusion detection with GA- based feature selection. *In: Proceedings of the 43rd annual southeast regional conference ACM*. 2005; 2:136–141p.
8. Wang X, Yang J et al. Feature selection based on rough sets and particle swarm optimization. *Pattern Recogn Lett*. 2007; 28: 459–471p.

9. Murphy, K.P. Naive Bayes classifiers: http://www.cs.ubc.ca/murphyk/Teaching/ CS340-Fall06/reading/NB.pdf
10. Han J., Kamber M., Pei J. *Data mining: concepts and techniques. Morgan kaufmann*. 2006.
11. Kazmierska Joanna, Julian Malicki. Application of the Naive Bayesian Classifier to optimize treatment decisions. *Radiotherapy and Oncology*. 2008; 86(2): 211–216p.
12. Kotsiantis Sotiris, Dimitris Kanellopoulos. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*. 2006; 32(1): 47–58p.
13. Muhlenbach Fabrice, Ricco Rakotomalala. Discretization of continuous attributes. *Encyclopedia of Data Warehousing and Mining*. 2005; 1: 397–402p.
14. Yang Ying, Geoffrey I. Webb. A comparative study of discretization methods for naive-bayes classifiers. *Proceedings of PKAW*. 2002.
15. Zibran, M. Chi-Squared test of independence.*University of Calgary, Canada. Retrieved from pages. cpsc. ucalgary. ca/~ saul/wiki/uploads/CPSC681/topic-fahim-CHI-Square. pdf*. 2012.
16. Easton Valerie J., John H. McColl. *Statistics glossary*. Steps. 1997.
17. Peaeson E., H. Haetlet. Biometrika tables for statisticians. *Biometrika Trust*. 1976.
18. Smith, Lisa F., Zandra S. Gratz, Suzanne G. Bousquet. *The art and practice of statistics*. CengageBrain. Com.2008.
19. Jeong D. H., Ziemkiewicz C., Ribarsky W., Chang R., Center C. V.Understanding Principal Component Analysis Using a Visual Analytics Tool. *Charlotte Visualization Center, UNC Charlotte. 2009.*
20. Abdi H, Williams LJ, Principal component analysis. *Statistics & data mining series, Wiley, New York*. 2010; 2: 433–459p.
21. Lakhina S, Joseph S, Verma B, Feature reduction using principal component analysis for effective anomaly-based intrusion detection on NSL–KDD. *International Journal of Engineering Science & Technology*. 2010; 2(6): 1790–1799p.
22. UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

Cite this Article

Biprodip Pal, Sadia Zaman, Md. Abu Hasan et al. Chi-Square Statistic and Principal Component Analysis Based Compressed Feature Selection Approach for Naïve Bayesian Classifier. *Journal of Artificial Intelligence Research & Advances*. 2015; 2(2): 16–23p.