Effective Approach for Classification of Nominal Data

Ketan Sanjay Desale
Assistant Professor
D.Y.Patil School of Engg. Academy, Ambi

Balaji Govind Shelale
Dept. of Computer Engineering,
D.Y.Patil School of Engg. Academy, Ambi

Sushant Navsare
Dept. of Computer Engineering,
D.Y.Patil School of Engg. Academy, Ambi

Dipak Bodade
Dept. of Computer Engineering,
D.Y.Patil School of Engg. Academy, Ambi

Krishnkumar.K.Khandelwal
Dept. of Computer Engineering,
D.Y.Patil School of Engg. Academy, Ambi

ABSTRACT
In today's era, network security has become very important and a severe issue in information and data security. The data present over the network is profoundly confidential. In order to perpetuate that data from malicious users a stable security framework is required. Intrusion detection system (IDS) is intended to detect illegitimate access to a computer or network systems. With advancement in technology by WWW, IDS can be the solution to stand guard the systems over the network. Over the time data mining techniques are used to develop efficient IDS. Here,a new approach is introduced by assembling data mining techniques such as data preprocessing, feature selection and classification for helping IDS to attain a higher detection rate. The proposed techniques have three building blocks: data preprocessing techniques are used to produce final subsets. Then, based on collected training subsets various feature selection methods are applied to remove irrelevant & redundant features. The efficiency of above ensemble is checked by applying it to the different classifiers such as naive bayes, J48. By experimental results, for credit-gdataset, using discretize or normalize filter with CAE accuracy of both classifiers i.e. naive bayes & J48 is increased. For vote dataset, using discretize or normalize filter with CFS accuracy of the naive bayes classifier increased.

General Terms
Classification, Feature Selection, Data Preprocessing.

Keywords
IDS, J48 classifier, Naive Bayes classifier.

1. INTRODUCTION
Current era is of networking & communication, due to the dramatic growth in the network computer resources, various network based applications have been developed which provides services to users in many different domain like E-commerce-Banking, public web services & defense services [1][2]. The advancement of network technology has been providing substantial enhancements for users generally. Paradoxically, new technologies bring new types of problems. The increase in the number of networked machines has led to an increase in unauthorized activity, not solely from external attackers, but conjointly from internal attackers, such as dissatisfied employees and people abusing their privileges for private gain. Peoples are keeping large amount of confidential data on networks, so we need strong security architecture to protect that data from attackers. IDS can be the solution to overcome the above problem. An intrusion detection system is a critical technology to detect intruders which are harmful to the system. Main goal of IDS is to protect system & network from intruders [3][4].IDS can detect many of the intrusions occurring in system but it generates too many false alerts & it also has a problem of low detection rate. The performance of IDS model can be increased using data mining techniques such as preprocessing, feature selection, classification.

Data preprocessing is often neglected but important step in data mining process. Data preprocessing technique which describes any sort of processing performed on raw information to prepare it for another analyzing procedure. Preprocessing reconstructs the data into a format that will be very easy and effective for further processing. There are various tools and techniques that are used for preprocessing which encompass: data cleaning, data integration, data reduction, data transformation and data discretization [5]. Data cleaning involves detecting & correcting the incorrect records. Data integration involves coupling data from various data sources. Data reduction is the process of abbreviating the amount of data that needs to be taken for mining process. In data transformation data is settled from one form to another which is appropriate for mining.

Feature selection (FS) is the process of extracting useful information from a largedataset. In FS duplicate valued and unnecessary features are discarded [6].FS is an effective approach which can be used to build efficient classification system. With reduced feature subset, the accuracy of classifier is improved [7].

Data classification is the method of organizing knowledge into classes for its simplest and economical use. It predicts catogorial class labels and classifies data to construct a model based on, training set and the values in a classifying attribute. There exist many classification techniques in data mining; in this paper only naive bayes and J48 classifier are used.

This paper is organized as follows: Section 2 provides related work based on the types of classifier design, the chosen feature selection method used for experiments. Section 3 provides the design of the proposed system. Section 4 and 5 contain experimental setup &results. Conclusion and discussion of future research is given at the end.

2. RELATED WORK

Feature selection is method which selectsessential subset of features according to some reasonable criterion so that original task can be achieved efficiently. By choosing anessential subset of features, insignificant and redundant features are removed according to criterion. Feature selection processes involve following steps: first is generation procedure which develop the next candidate subset; second is an evaluation function which evaluates the subset and third is a stopping criterion to determine when to stop; and last step is a validation procedure to check for the validation of dataset [8] [9].Here, 3 feature selection methods are used for dataset evaluation. The methods are: CFS, CAE& IG.

2.1 Correlation Feature Selection (CFS)

Correlation featureselection (CFS) is a heuristic way to evaluate the value of a features subset. A good feature subset is a subset that has features which are highly associated (predictive of) with the class, yet unassociated (not predictive of) with each other.CFS measures relations betweennominal features, so numeric features are discretized first. However, the concept of correlation-based feature selection does not rely on any particular data transformation [10].A function that calculates the best individual feature is given by:

$$HM = \frac{n * r_{fc}}{\sqrt{n + n * (n-1) * r_{ff}}}$$

Where, HMis the heuristic merit of a features subset *S*containing *n features*, r_{fc} is the average (avg.) feature-class correlation, and r_{ff} is the avg. feature-feature correlation.

In above equation, numerator points to how predictive a group of features are; and the denominator points to how much redundancy there is among those features.

2.2 CAE

The classification methods were designed to minimize the errors. Real world applications requires classifiers to reduce overall cost, which involves false classification cost (every error has associated cost) and attribute cost. CAE also called as cost-sensitive classification. The main aim of using CAE is to reduce cost of the classification.

Cost function:

$$\frac{\Delta A_j}{(B_j \emptyset_j)^{\omega}}$$

Where, ΔA_j is gain ratio for attribute j, B_j is cost of attribute j, $\emptyset_j$ is risk element related with attribute j and ω is scale factor for cost.

2.3 Information Gain (IG)

Information Gain guides us to determine which feature of the class is most useful for classification, using its entropy value. Entropy is indicated by the information content of a feature or how much information that features is giving us. More the information content, the higher the entropy, IG value is calculated as:

IG (T, v) = E (T)–E (T| v)

Where Eis the information entropy, Tis a training example, and vis a variable value. Above equation, calculates the IG values of that a training example T obtains from an observation that a random variable A takes some value v.

In this paper, two classifiers i.e. naive bayes and J48 classifier are used for comparison. Comparison is made on accuracy.

2.4 J48 Classifier Algorithm

Depending on the attribute values, it creates a decision tree. The decision tree approach is most helpful in classification problem. With this system, a tree is built to model the classification method. Once the tree is built, it's applied to every tuple within the database which results in classification for that record. While building a decision tree, J48 ignores the omitted values. J48 allows classification based on decision trees or rules generated from that decision tree. [11][13].

INPUT :

TD //Training data

OUTPUT :

T //Decision tree

DTBUILD (*TD)

{

T=φ;

T= Create root node and label with splitting attribute;

T= Add arc to root node for each split predicate and

Label;

For each arc do

TD= Database created by applying splitting

predicate to TD;

If stopping point reached for this path, then

T'= create leaf node and label with

appropriate class;

Else

T'= DTBUILD (TD);

T= add T' to arc;

}

2.5 Naive Bayes

A naive bayes classifier is a probabilistic classifier which implements the bayes theorem with a naive (strong) assumption. Assumption is features that describe the objects which are to be classified are analytically independent from each other. In spite of this assumption naive bayes is very effective in real world application [12].

The Bayes Theorem:

$$P(H \mid Z) = \frac{P(Z \mid H)\ P(H)}{P(Z)}$$

$P(H \mid Z)$: Posterior Probability of H

$P(Z \mid H)$: Posterior Probability of Z

$P(H)$: Prior Probability of H

(Z): Prior Probability of Z

3. PROPOSED SYSTEM

The product of preprocessing is final training sets. The overall system architecture of the proposed approach is described in Figure 1. Proposed architecture consists of dataset, data pre-

processing techniques, feature selection methods & classifiers. This paper describes an effective approach for classification of nominal data. For this experiment credit-g & vote dataset is used. Credit-g dataset with 1000 instances & vote dataset with 435 instances are used for training purpose.

Step 1: Data Preprocessing:

In first step data is preprocessed by applying various pre-processing techniques out of which discretize,normalize & principal components are used. In this step, input values are sorted and then divided into n blocks called as bins. Then all values in each bin are mapped to nominal type.

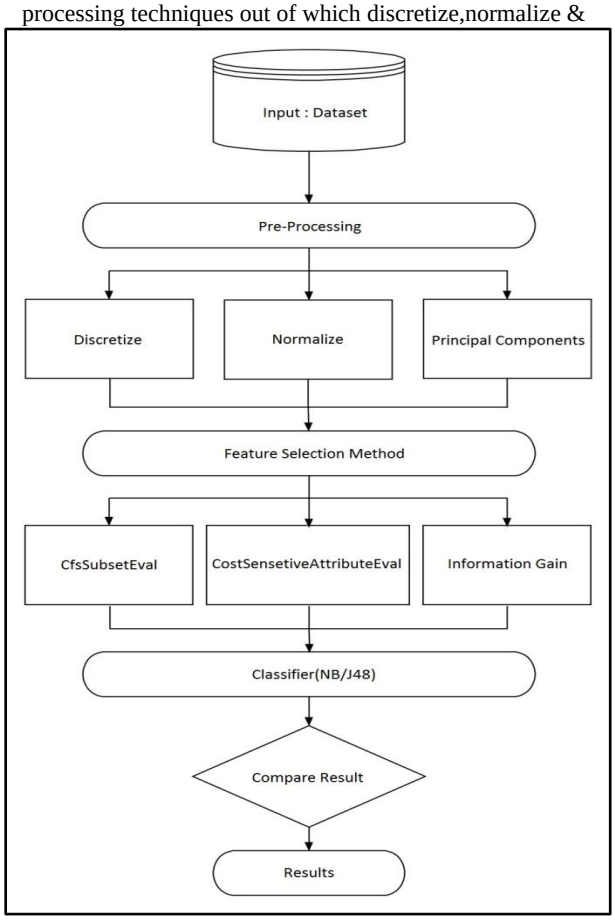

Fig 1: System Architecture

Step 2: Feature selection

In this step, feature selection methods are applied to the processed data. There are various feature selection methods;Here,CfsSubsetEval(CFS), CostSensitiveAttributeEval(CAE) & Information Gain (IG) methods are used.

Step 3: Classification

Finally, forperformance testing, features selected from above step are applied to naive bayes and J48 classifiers. The obtained results are then compared using parameter like accuracy.

4. EXPERIMENTAL SETUP

WEKA is a Software tool that contains a large collection of machine learning algorithms implemented in Java. The main types of learning problems that it can handle: classification, regression, attribute selection methods, clustering. It also has some support for association rule mining. WEKA has three built-in graphical user interfaces:

- Explorer: It is the most popular interface for Data Processing.
- Knowledge Flow: It helps to visualize flow of Data.
- Experimenter: It is a framework where experiments are performed which enables large scale statistical comparison between learning algorithms.

It is also possible to operate WEKA from command line interface. All the experiments are performed on the system having configuration of 2.40 GHz i5 Processor, 4GB RAM &1TB hard drive.

4.1 Dataset Used

Dataset consists of all the information collected during a survey that needs to be analyzed. Here dataset credit-g and vote are used for evaluation.

Following table shows the detailed description of datasets which are used for experiments.

Table 1: Dataset Used for Experiment

Dataset used	Instances
Credit-g	1000
Vote	435

5. RESULTS AND DISCUSSION

The effectiveness of the proposed technique is evaluated by performing experiments using WEKA tool with data preprocessing technique, feature selection methods & classifiers. Following sections shows the results of evaluation using the WEKA tool in graphical form.

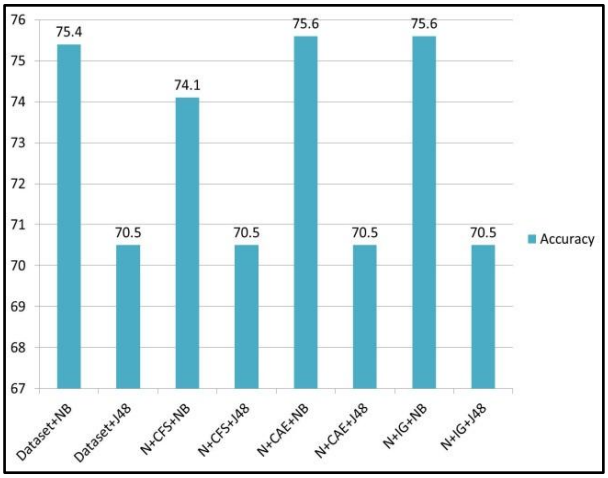

Fig 2: DatasetCredit-G with Normalization

Figure 2 shows dataset credit-g is selected for evaluation of results. Without proposed approach, dataset was applied to naive bayes classifier, obtained accuracy was 75.4% whereas with proposed approach using normalize preprocessing & CAE, IG feature selection method, accuracy was increased considerably to 75.60%. For CFS feature selection method accuracy is decreased. Without proposed approach, dataset was applied to J48 classifier, obtained accuracy was 70.5%. While with proposed approach obtained accuracy was same.

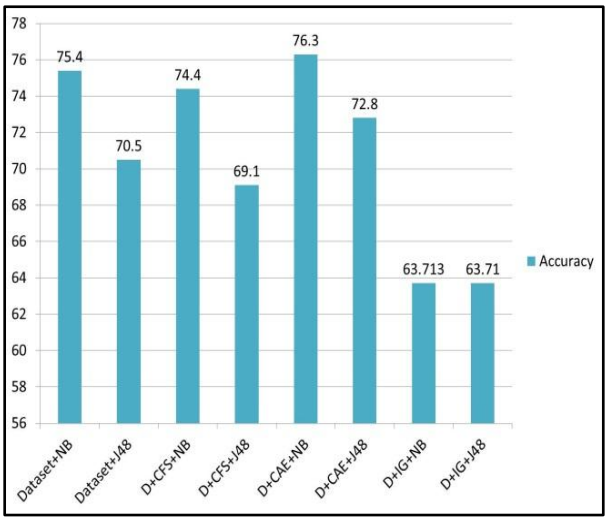

Fig 3: Dataset Credit-G with Discretize

The figure 3 shows dataset credit-g was selected for evaluation of results. Without proposed approach, dataset was applied with naive bayes classifier, obtained accuracy was 75.4% whereas with a proposed approach using discretize preprocessing & CAE feature selection method, accuracy was increased considerably to 76.30%. For others feature selection methods accuracy decreased considerably. Without proposed approach, dataset was applied to J48classifier, obtained accuracy was 70.5%. Whereas with a proposed approach using discretize preprocessing & CAE feature selection method, accuracy was increased considerably to 72.80%, for other feature selection methods accuracy was decreased.

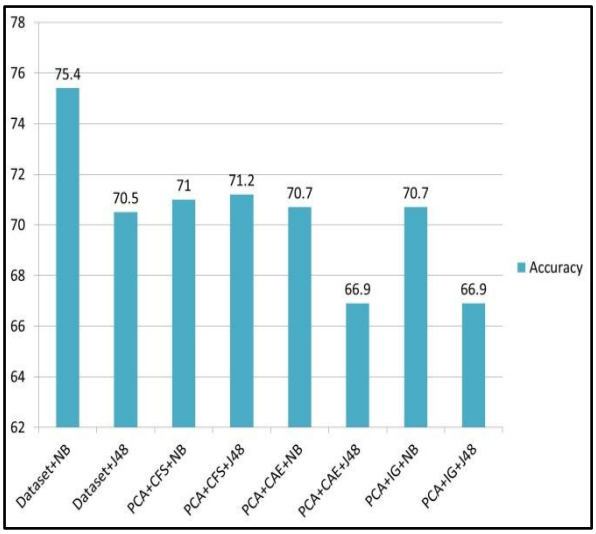

Fig 4: Dataset Credit-G with PCA

The figure4showsdataset credit-g was selected for evaluation of results. Without proposed approach, dataset was applied with naive bayes classifier, obtained accuracy was 75.4% whereas with a proposed approach using PCA preprocessing & other feature selection methods accuracy decreased. Without proposed approach, dataset was applied to J48 classifier, obtained accuracy was 70.5%.Whereas with a proposed approach using discretize preprocessing & CFS feature selection method, accuracy was increased considerably to 71.20%, for other feature selection methods accuracy was decreased.

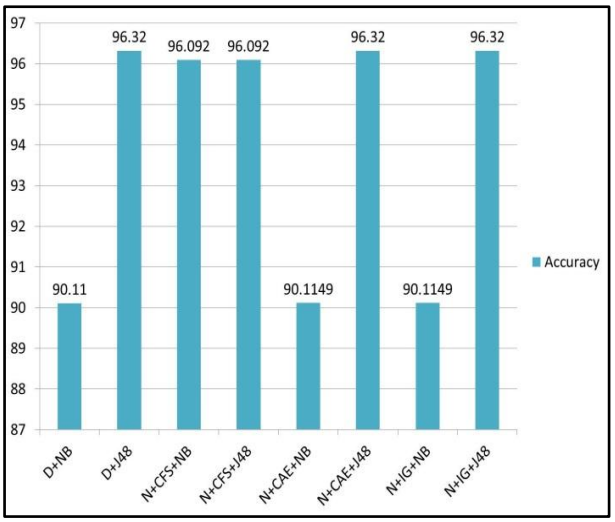

Fig 5: Dataset Vote with Normalize

Figure 5 shows dataset vote was selected for evaluation of results. Without proposed approach, dataset was applied to naive bayes classifier, obtained accuracy was 90.11%. Whereas with a proposed approach using normalized preprocessing & CFS feature selection method, accuracy was increased considerably to 96.092%, for other feature selection methods accuracy is same. Without proposed approach, dataset is applied to J48 classifier, obtained accuracy was 96.32%. Whereas, with proposed approach using normalizepreprocessing & CAE, CFS, IG accuracy was not affected much.

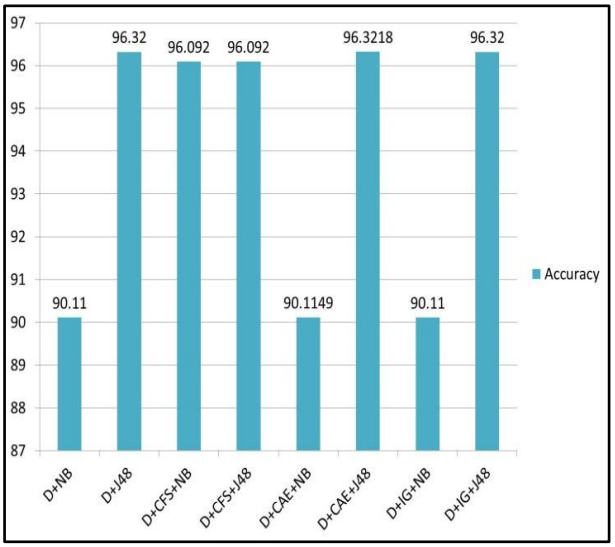

Fig 6: Dataset Vote with Discretize

Figure 6 shows dataset vote was selected for evaluation of results. Without proposed approach, dataset was applied to naive bayes classifier, obtained accuracy was 90.11%.whereaswith a proposed approach using discretize preprocessing & CFS feature selection method, accuracy was increased considerably to 96.092%, for other feature selection methods accuracy was same. Without proposed approach, dataset was applied to J48 classifier, obtained accuracy was 96.32%.Whereas, with a proposed approach using discretize preprocessing & CAE, CFS, IG feature selection methods, accuracy not affected much.

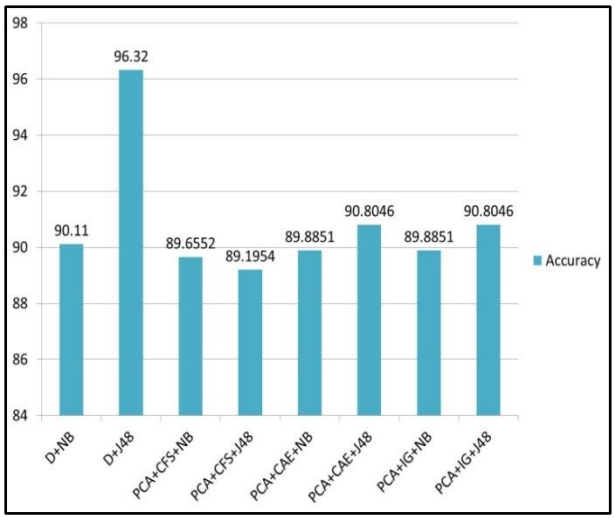

Fig7: Dataset Vote with PCA

Figure 7 shows dataset vote was selected for evaluation of results. Without proposed approach, dataset was applied to naive bayes classifier, obtained accuracy was 90.11%. Whereas with a proposed approach using PCA preprocessing & other feature selection methods, accuracy was decreased considerably to 89.6552%. Without proposed approach, dataset was applied to J48 classifier, obtained accuracy was 96.32%, whereas with a proposed approach using PCA preprocessing & CAE, CFS, IG accuracy is decreased to 89.8851%.

6. CONCLUSION

In this paper, the effective approach for selecting classification technique for nominal data was discussed. Data preprocessing is done using discretize, normalize & PCA, further feature selection is done using CAE, CFS& IG. Then the effect of the proposed technique is investigated using two classifiers i.e. naive bayes and J48.

From experimental results, it can be concluded that, for credit-gdataset using discretize or normalize filter with CAE accuracy of both classifiers i.e. naive bayes & J48 is increased. While using PCA, accuracy doesn't influence much.

For vote dataset, using normalize & discretize filter with a CFS accuracy of the naive bayes classifier increased terrifically, but for J48 accuracy doesn't excite much. In applications where accuracy has a major concern, one can use above mentioned proposed approach.

Future work would be focused on applying the proposed approach on other data types like numeric data types, characters & multidimensional data.

7. ACKNOWLEDGMENT

We are thankful to facultyof Computer Engineering Department, Savitribai Phule Pune University for their support. The product of this research paper would not be possible without all of them.

8. REFERENCES

[1] J. Gomez & D. Dasgupta, (2002), S. K., and Peterson, L. L. 1993. Reasoning about naming systems.

[2] Mr. Suraj S. Morkhade1, Prof. Mahip Bartere2,"Survey on Data Mining based Intrusion Detection Systems", International Journal of Application or Innovation in Engineering & Management (IJAIEM)

[3] J. Gomez & D. Dasgupta, (2002) "Evolving Fuzzy Classifiers for Intrusion Detection", IEEEProceedings of the IEEE Workshop on Information Assurance, West Point, NY.

[4] R. H. Gong, M. Zulkernine & P. Abolmaesumi, (2005) "A Software Implementation of a GeneticAlgorithm Based Approach to Network Intrusion Detection", Sixth InternationalConference on Software Engineering, Artificial Intelligence, Networking and Parallel/DistributedComputing and First ACIS International Workshop on Self-Assembling Wireless Networks.

[5] Jiawei Han,Micheline Kamber,Jian Pei,"Data Mining : Concept and Techniques ", 3rd edition, Morgan Kaufmann,2011.(1st edition.,2000-2001)(2nd edition 2006).

[6] H Liu andL Yu "Feature Selection for High-Dimensional Data – A Fast Correlation-Based Filter Solution", In Machine Learning-International Workshop Then Conference, 2003, Vol. 20(2), p. 856

[7] P Langley, Selection of Relevant Features in Machine Learning, DefenseTechnical Information Center, 1994, pp. 140-144.

[8] J Hua, WD Tembe, ER Dougherty, Performance of feature-selection methods in the classification of high-dimension data, Pattern Recognition, 2009, Vol. 42(3), pp. 409-424.

[9] H Liu, H Motoda, L Yu,Feature selection with selective sampling,Machine Learning-International Workshop Then Conference, 2002, pp.395-402.

[10] Mark A. Hall, Lloyd A. Smith , "Practical Feature Subset Selection forMachine Learning", Computer ScienceDepartment, University of Waikato, Hamilton, New Zealand.

[11] Margaret H. Danham,S. Sridhar, " Data mining,Introductory and Advanced Topics", Personeducation , 1st ed., 2006

[12] George Dimitoglou, James A. Adams, and CarolM. Jim," Comparison of the C4.5 and a Naive BayesClassifier for the Prediction of Lung CancerSurvivability"

[13] Aman Kumar Sharma, Suruchi Sahni, "AComparative Study of Classification Algorithms forSpam Email Data Analysis", IJCSE, Vol. 3, No. 5, 2011, pp. 1890-1895