Sequence Classifier Methods on Numerous Item set Mining by SVM in Operation Dataset

K. Sivakumar A.S. Prakaash

Abstract: ***The adoption tree of the number of item sets includes the most recapitulate form of method unreservedly or apparently. As for the example, Support Vector Machine (SVM) system can be used to find out the approximate tree in the span of fashion with the accompany of the age group. Constant systems, definite or indefinite, helps to evaluate the estimation of the tree of item sets. In the sequence to explore the classification tree in the wide shape with the collaboration aspirant based generation SVM system can be executed. On the basis of classification tree is formed upon the joining, the execution tree would be used for the garnishing of the candidate. Systems like SVM take the help of vertical relationships among the prognosis databases in order to avoid the re-doing of the computation of the work done for the shorter designs comprising different dimensions. With the usage of maximum and closed designs of quick styles of the prospecting methods much better performances can also be attained. Full of effectiveness-based development for quick figures prospecting systems are also analyzed in this blog. This paper most importantly reviewed about the quick designs of prospecting systems.***

I. INTRODUCTION

Consecutive phenomena or objects are found in most of the actual datasets, such as variety of texts, speech, indications; videos web usage logbooks and anatomical formation. Constant categorization has always been an important topic in data tapping and in statistical tool learning because of important trouble in statistical learning and archives prospecting because of a wide range of supplication in use. Constant enumeration is known as the appropriate labels to new consecutive based on the skills secure in the build-up period. Constant classification is known as the lot of experimentation in the place of amalgamating pattern of prospecting procedures and enumeration such as enumeration based on equate rules, continuous designs based on constant classifiers and many others. Amalgamation of all these mechanisms can cultivate better outcomes. The main goal of this paper is to attain better conclusion. The main focus of this paper is to acquire constant classifier by using a novel piece set tapping skill. The set of a piece should be constant depending on the closeness of the circumstance of each piece and how repeatedly the item occurs. So this paper

comes up with a new procedure called constant enumeration on item sets. A new method has been proposed by this paper which is termed as the sequence classification with respect to various sets of items that are not specifically taken into consideration. For achieving a strong reduction in terms of various complications in comparison with the sequence classifiers, it will produce and analyze the certain patterns by using the connectivity that can take this location of sequences into account. There has been a valued introduction of two classifiers of statistics, first by utilizing the cohesion in order to execute the sequential character of the data into certain methods and secondly by using such sets of items in order to ignore the various complicacies that are associated with the extracting of the sequential pattern which in turn denote the fact that both the methods can provide the reasonable results.

II. RELATED WORK

Richaruna Taroz and Alminat Trioup [1] stated that capturing the various types of work with regard to the types of decision tree learning which requires some tuning. However the SVMs and the neural nets need careful selection of different parameters. Their valuable research and analysis on STATLOG data set declares the fact that the method of learning includes various aspects namely Bagging, Random Forests, SVMs & boosting in order to acquire the appropriate style of performance which could have been very much strenuous to attain 15 years ago. SVM is basically known for Support Vector Machine and integer-coded Genetic Algorithm (GA) which is highly used for detecting the classification of heart disease by Suman Sexena, Prakash Pandey, and B.N. Namkar [2]. Y. Hrusta, D. Patil, and A. Rajput [3], have stated that Support Vector Machine and RBF are rendering high accuracy in various purpose of classification. Nigam Trenak [4] in his research study briefly analyzed that the educational mining concentrates on evaluating the performances of the student rather than evaluating the performances of the instructor. The common techniques have been identified to analyze the performances of the instructor in the due intervention of detecting the various questionnaires to identify the perception of the students. This paper focuses on four various techniques of classification that is – support vector machines, algorithm on decision tree, artificial neural networks which are to be used to build up the various models of classifications. Their activities are then distinguished with respect to composed set of data in connection with the responses of the students for evaluating the real course questionnaire linked with precision, performance metrics. Though various models of classifier, it showcase the high level of performance classification, the popularly known C5.0 classifier act as the best in terms of accuracy and specificity. Therefore, the analysis of the variability of each and every model of classifier is successfully done. Now the analysis shows that various questions in identifying the questionnaire seem to be unrelated.

Revised Manuscript Received on December 30, 2018.

K. Sivakumar, Professor, Department of Mathematics, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India.

A.S. Prakaash, Research Scholar, Department of Mathematics, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India.

The analysis also revealed that result that the progress of the instructor is generally based upon the perception of the student.

S.K.Nautaiya, Jimi Rathore [5] in his research analysis discovered the fact that Generalized Sequential Pattern uses certain algorithm in order to search the relevant patterns from the database record of various students. The tree based frequent pattern can be used for the purpose of detecting performance of the student whether they are failing or passing. However, when it is identified that the student is at the risk zone of failure, they needs to be provided with the guidance of improving their performances.

The interesting rule has been discovered between the various sets of attributes in the datasets framed by the Association-rule. The association rule is represented by various datasets along with their inter-relationship. Such sort of information are generally used for the purpose of undertaking certain strategic decisions including wide variety of tasks that includes shelf management, advancement in pricing and much more[7]. The rule of Traditional Association mainly includes data analyst who are assigned with the various company's datasets with the aim of finding the rules of pattern which is present between the various datasets [8]. However we can acquire various sophisticated analysis which are based upon the large data sets in a more cost-effective style. It also involves some sort of security risk on behalf of the proprietor of the data where the delicate information can be removed by the data miner. However in the present times, the KDD is still considered to be the widely used technique for mining of the association rule.

III. OBJECTIVES

The usage of World Wide Web has been increased to such an extent that large amount of data can be automatically enhanced by the web servers. By analyzing the significant access to the data server, you can obtain useful information. The issue of web usage mining from two or more web servers with the purpose of detecting the relationship between the stored data by investing specific concentration to the new patterns of the data which is depicted in this paper. This is how the statistical rule of algorithm are efficiently implemented for matching with the new patterns in order to calculate the various measures of the new frequent pattern in the specific context.

IV. DATA PREPROCESSING

There are various steps included in the data preprocessing. Some of these steps are mentioned below:

1. The first step is cleaning of data,
2. The second step is integration of data
3. The third step is reduction of data and
4. The fourth step is transformation of data.

Cleaning of data-The data cleaning is the first step which is involved in the preprocessing of data. It aims at cleaning the data by detecting & removing the deviation and resolving certain obstacles and inconsistency. If the users somehow keep the view that all the data are inaccurate, than it is obvious that the users will not believe the result of the data mining that has been sleeked. Filthy issues of data or any sort of inconsistencies in the procedure of data mining can result into unpredictable output. It has been observed that most of the mining posses some sort of techniques or methods in order to cope up with the incomplete data. For the purpose of achieving that, one should focus on avoiding the over fitting of data. Thus the helpful process of running the prepossessing of data needs adequate routine of data cleaning. Data integration simply refers to the process of combining multiple databases, files and data cubes. Combination of various data using certain technologies tends to assure a unified source and view of data. It becomes highly important in certain cases where two companies merging system or the consolidated applications in one company gives a unified sights of the data assets of the company. Implementing the data integration is popularly well known for building the data warehouse for the enterprise which in turn enables a business to analyze certain statistical tools in the data warehouse because the source system does not involve any sort of corresponding data although such data are named identically which may be called to different entities. Data integration can prove to be a real changing situation when the attributes of the same kind of data posses different identical names in the databases which can seriously cause instability and dismissal. Therefore, the process of cleaning and integration of data performs as a most important step for data preprocessing while creating the data for warehouse. Data cleaning also helps in identifying and removing the dismissals originating from the data integration process.

Reduction of data or data reduction ensures the representation of the reduced data set that is very much little in terms of its volume yet introducing the same analyzed results. People wonder that if the selected data for the purpose of concrete analysis is huge, then it will definitely lower the process of data mining. There should be some way of reducing the data size without disturbing the data mining results. Data reduction as per dimensionality and numerosity are considered to be the sound strategies that need to be included. Dimensionality reduction of data is basically done with the help of the compressed description of the real data with the application of the encoding data schemes. It includes attribute subset selection of attributes, techniques representing the data compression and construction of data attributes. Data replacement is usually done by using various alternatives, small size representation of data with the help of certain parametric and non parametric models in case of certain data reduction category.

Transformation of data is the fourth step which is involved in the data preprocessing. It is the process of conversion of data from one layout/ format to another structure or format. It is considered as the basic aspects of data integration, data warehousing, data cleaning and integrating various applications. It uses mining based algorithm for the purpose of analyzing various aspects which includes closest classifiers, neural networks or clustering.

The analyzed data is normalized to showcase the proven results by scaling the data into smaller range that is [0.0, 1.0]. The process of data transformation mainly includes discovery of data, data mapping, generation of data code, execution of code and the data review.

The concept of data hierarchy generation, normalization and discretization are some of the few forms of transforming data where as in case of customer data, it contains the age and annual salary attributes. The value of the attributes that comprises the annual salary is usually larger than the age attributes. So the measurement of distance based on the annual salary attributes if left unnormalized, than it may generally exceed the measurement of distance which is taken on the basis of age attributes.

V. METHODOLOGY SVM

SVM Algorithms

The SVM version is easy and fixed

Notation

The succeeding footnote is used all through in this segment unless otherwise stated:

j0 Total number of forecaster.

X Categorial forecaster point X' =(X1, …,Xj),where j is the number of forecaster survey

Mj Number of classification for predictor Xj.

Y Categorial objective variable.

K Number of classification of Y

N Total number of instances or styles in the training input.

Nk The number of instances with Y= k in the training input.

Njmk The number of instances with Y= k and Xj=m in the training input. . πk The possibility For Y=k

pjmk The possibility of Xj=m given Y=k

SVM Model

It is based on the conditional autonomy model of each forecaster given the target class. By Bayes' theorem, the after probability of Y given X is:

$$P(Y=k|\mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x}|Y=k)P(Y=k)}{\sum_{i=1}^{K} P(\mathbf{X}=\mathbf{x}|Y=i)\,P(Y=i)}$$

Let X1, ...,Xj be the j predictors considered in the model. The SVM model understands that X1, ...,Xj are conditionally independent given the target; that is:

$$P\left(\mathbf{X}=\mathbf{x}\,|Y=k\right) = \prod_{j=1}^{J} P\left(X_j=x_j|\,Y=k\right)$$

These possibilities are calculated from training input by the following equations:

$$\pi_k = P(Y=k) = \frac{N_k+\lambda}{N+K\lambda}$$

$$p_{mk}^{j} = P(X_j=m|Y=k) = \frac{N_{mk}^{j}+f}{\sum_{l=1}^{M_j} N_{lk}^{j} + M_j f}$$

SVM Algorithms

Where N_k is evaluated on the basis of all non-missing Y, where as Nj, mk is based on all non-missing pairs of Xj and Y, and the factors λ and f are established to solve the problems caused by zero or very small cell counts. These calculations correlate to Bayesian calculation of the series possibilities with Dirichlet priors. Factual studies suggest (Kohavi et al., 1997).

A single input pass is needed to gather all the required counts.

For the special condition in which j = 0; that is, there is no forecaster at all. When there would be vacant classifications in the target variable or classified predictors, these vacant categories should be eliminated from the calculations.

Preprocessing

Missing Values

If every benefit is missing or if it has only one observation category the forecaster is avoided. A case is avoided if the benefit of the target variable or the benefit of all the forecasters are missing. For each case misplaced some, but not all, of the values of the forecasters, only the forecasters with non-missing values are used to forecast the case, as suggested in (Kohavi et al., 1997).

This implies the following equation:

$$P(\mathbf{X}=\mathbf{x}_i|Y=y_i) = \prod_{\{j:x_{ji}\text{not missing}\}} P(X_j=x_{ji}|Y=y_i)$$

This also implies the following equation for B(j) in average log-likelihood calculations:

$$B(J) = -\frac{1}{N'}\sum_{i=1}^{N} \log\left(\sum_{k=1}^{K}\pi_k \prod_{\{j:x_{ji}\text{not missing}\}} p_{x_{ji}k}^{j}\right)$$

Where the log() term for the case i is avoided if all the values of the predictors examine in the model ar missing, and N' is the total number of terms that are not avoided in the sum

Continuous Variables

The SVM model ascertains the target and the predictor variables are classified. If there are constant variables, they need to be separated. The easiest way to separate a continuous variable is by dividing the estate of a variable into equal span bins. This technique performs well with the SVM model while no obvious up gradation is found when the complex procedures are used (Hsu et al., 2000).

Equal frequency binning techniques may construct vacant bins. In this case, vacant bins are removed by changing bin frontier points. Let b1 < b2< …bn be the bin frontier points produced by the equal Sometimes the span binning technique. The two end bins [-∞, b1] and (bn, ∞) are non-vacant by deigns. Suppose that bin (bi, bi+1] is vacant, and suppose that the closest left SVM.

Algorithms Steps

Non-vacant bin has right frontier point bj (<bi)and the closest right non-vacant box has left point bk (> bi).

Then the vacant boxes are eliminated by deleting all the frontier points from bj to bk and setting anew frontier point at (bj+bk)/2.

Feature Selection

Given a total of j0 forecasters, the goal of feature selection is to choose a portion of j forecasters using SVM model (natarajan and Pednault, 2001). This technique has the following steps:

Collaborating the necessary summary statistics to evaluate all the possible model guidelines. Create a series of candidate forecaster group that has an increasing number of forecasters; that is; each serial subset is equal to the previous subset plus one more forecaster." The "best" subset can be found from this series.

Collect Summary Statistics

One pass through the training input is required to collaborate the total number of cases, the number of cases per classification of the target variable, and the number of cases per classification of the target variable for each group for each forecaster.

Create the Sequence of Subsets

Starting the group of forecasters considered important to the model, which can be vacant and can be used to start with. For each forecaster it is not in the group, a SVM model is perfect with the forecaster plus the forecasters in the group. The predictor that gives the huge log-likelihood is added to create the next larger group. This repeats until the model includes the user-specified:

Accurate number of predictors or Maximum Number of forecasters

Alternatively, the maximum number of forecasters, Max, may be automatically chosen by the following equation:

$$J_{\text{Max}} = \min\left\{J_{\text{Must}} + \min\left\{100, \max\left(20, \frac{J_0}{5}\right)\right\}, J_0\right\}$$

Where j must is the number of predictors in the beginning subset.

VI. SVM Algorithms

Find the "Best" Subset

If you particularize an accurate number of predictors, the annual group in the series is the annual model. If you particularize a maximum number of predictors, the "best'' group is regulates by one of the following:

A test input standard based on the average log-likelihood of the test input.

1. A pseudo-BIC standard that uses the average log-likelihood of the training input and punishes overly complex models using the number of forecasters and the number of cases. This criteria is used if there are no test input.
2. A pseudo-BIC criterion that uses the average log-likelihood of the training data and penalizes overly complex models using the number of predictors and number of cases. This criterion is used if there are no test data

Smaller values of these criteria indicate "better" models. The "best" group is the one with the smallest value of the criterion used.

Test Data Criterion

$$Q(J) = -\bar{l}_{\text{Test}}(J)$$

Where Test is the average log-likelihood for test data.

Pseudo-BIC Criterion

$$Q(J) = -\bar{l}_{\text{Train}}(J) + \frac{1}{2} J \frac{\log(N)}{N}$$

Where j indicates the number of predictors available in the model, and Train indicates the average log-likelihood for training data.

Average Log-Likelihood

The mean (conditional) log-likelihood for data with j predictors is

$$\begin{aligned}\bar{l}(J) &= \frac{1}{N}\log L = \frac{1}{N}\sum_{i=1}^{N}\log P(Y=y_i|\mathbf{X}=\mathbf{x}_i)\\ &= \frac{1}{N}\sum_{i=1}^{N}\log P(Y=y_i) + \frac{1}{N}\sum_{i=1}^{N}\log P(\mathbf{X}=\mathbf{x}_i|Y=y_i)\\ &\quad - \frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}P(\mathbf{X}=\mathbf{x}_i|Y=k)P(Y=k)\right)\\ &= \frac{1}{N}\sum_{k=1}^{K}N_k\log(\pi_k) + \frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{m=1}^{M_j}N_{mk}^{j}\log\left(p_{mk}^{j}\right) - \frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}\pi_k\prod_{j=1}^{J}p_{x_{ji}k}^{j}\right)\end{aligned}$$

Let

$$A(J) = \frac{1}{N}\sum_{k=1}^{K}N_k\log(\pi_k) + \frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{m=1}^{M_j}N_{mk}^{j}\log\left(p_{mk}^{j}\right)$$

SVM Algorithms

$$B(J) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}\pi_k\prod_{j=1}^{J}p_{x_{ji}k}^{j}\right)$$

then

$$\bar{l}(J) = A(J) + B(J)$$

Note: To calculate the exceptional situation in which j = 0; that is, there are no predictors,

$$\bar{l}(J) = \frac{1}{N}\sum_{i=1}^{N}\log P(Y=y_i) = \frac{1}{N}\sum_{k=1}^{K}N_k\log(\pi_k)$$

Calculation of Average Log-Likelihood by Sampling

When there is a need to add each and every predictor to maintain the subset sequence, then a data pass will be required for calculating the B(j). It is not a major issue if the data set is little small to perfectly fit in the memory. If it does not get fit into the memory, then it can prove to be a costly affair. In order to calculate the B(j), the SVM generally uses replicate data. As per the various research results, it has been showcased that this method yields sound results.

The B(j) formula can be rewritten for m cases data set as:

$$B(J) = -\frac{1}{m}\sum_{i=1}^{m}\log\left(\sum_{k=1}^{K}\pi_k\prod_{j=1}^{J}p^{j}_{x_{ji}k}\right)$$

By default m = 1000.

Classification

The target category with the highest posterior probability is the predicted category for a given case.

$$\hat{y}(\mathbf{x}) = \underset{k}{\operatorname{argmax}}\{P(Y=k|\mathbf{X}=\mathbf{x})\} = \underset{k}{\operatorname{argmax}}\{P(\mathbf{X}=\mathbf{x}|Y=k)P(Y=k)\}$$

Ties that are created in relation with the target category are served with greater prior probability πk. In certain cases when the categorical predictors do not take place in the training data, then such categories will be considered as missing.

Classification Error

The ratio of miscalculation on test data represents equal errors due to the presence of any test data. In the absence of the test data, the error will be equal to the misclassification ratio of the training data.

VII. CONCLUSION

Due to the availability of plenty of applications in various data mining problems like classification and clustering, the issue of frequent pattern mining was broadly deliberated in the literature. The data processing includes certain major steps namely data reduction process, integration of data and transformation of data which in turn explains about the usages of various data. Various domains that are diverse, posses the detection of software bug along with the applications which are being employed to detect the pattern of frequent mining. It has been broadly explored in the various aspects of algorithm. Therefore, in this study it has been mentioned that in what way the present frequent mining data issues has been looked out with respect to the SVM model usage.

REFERENCES

1. Rich Caruana and Alexandru Niculescu-Mizil (2006), An Empirical Comparison of Supervised Learning Algorithm, 23rd International Conference on Machine Learning, Pittsburgh.
2. S.Bhatia, P. Prakash and G.N. Pillai, SVM based Decision Support System for Heart Disease Classification with Integercoded Genetic Algorithm to select critical features, Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, pp.34-38, 2008.
3. S.Ghumbre, C. Patil, and A.Ghatol, Heart disease diagnosis using support vector machine,Proceedings of the International Conference on Computer Science and Information Technology (ICCSIT '11), Pattaya, Thailand, 2011.
4. Priyanka AnandraoPatil, R. V. Mane," Prediction of Students Performance Using Frequent Pattern Tree," Sixth International Conference on Computational Intelligence and Communication Networks,2014
5. Mustafa Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education," IEEE Access , Volume: 4 ,2016.
6. P Deepa Shenoy, Srinivasa K G, Venugopal K R and L M Patnaik, "Dynamic Association Rule Mining using Genetic Algorithms," Intelligent Data Analysis, vol. 9, no. 5, pp. 439–453, 2005.
7. P Deepa Shenoy, Srinivasa K G, Venugopal K R and L M Patnaik,"Evolutionary Approach for Mining Association Rules on Dynamic Databases,"7th Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD)2003, Seoul, South Korea, pp. 325–336, 2003.
8. S. J. Rizvi and J. R. Haritsa, "Maintaining Data Privacy in Association Rule Mining," Proceedings of the 28th International Conference on Very Large Data Bases, pp. 682–693, 2002.
9. S. M. Darwish, M. M. Madbouly, and M. A. ElHakeem, "A Database Sanitizing Algorithm for Hiding Sensitive Multi-level Association Rule mining," International Journal of Computer and Communication Engineering, vol. 3, no. 4, pp. 285– 293, 2014.
10. J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639–644, 2002.
11. J. Vaidya and C. Clifton, "Secure Set Intersection Cardinality with Application to Association Rule Mining," Journal of Computer Security, vol. 13, no. 4, pp. 593–622, 2005.
12. Kaur, P., Kaushal, S Security concerns in cloud computing. In: Accepted For International Conference on High Performance Architecture And Grid Computing-2011. Chitkara University, Rajpura (2011)