International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 6, November-December 2014 ISSN 2278-6856

A Survey on Dimensionality Reduction Technique

V. Arul Kumar[1], N. Elavarasan[2],

[1]Assistant Professor in Computer Applications, Thanthai Hans Roever College, Perambalur.

[2]Research Scholar, Nehru Memorial College (Autonomous), Puthanampatti.

Abstract

Data mining is an automatic extraction of useful, often previously unknown information from large databases or data sets. The data collected from the real world applications contain lots of erroneous data. Data preprocessing is an important technique in data mining to rectify the erroneous data present in the dataset. Many data mining applications contain high dimensional data. The High dimensionality decreases the performance of the mining algorithms and increases the time and space required for processing the data. The high dimensionality issue is resolved using the Dimensionality Reduction (DR) technique. The DR is divided into two: feature selection and feature extraction. In this paper a detail survey has been carried out to know how the dimensionality problem has solved by using the two different techniques. And also various statistical measures are explained to select the most relevant features and different statistical techniques are analysed to extract the new set of features form the original features

Keywords:- Dimensionality Reduction, Feature Selection, Feature Extraction, Principal Component Analysis, Principal Feature Analysis, Linear Discriminant Analysis.

1. INTRODUCTION

In recent years, lot of improvement in the database technology leads to generation of large volume of digital data. The increase of data size becomes a great challenging task for the humans to discover or to extract a valuable information. With the help of data mining technique this task became an easiest one. Data Mining is the process of searching valuable information in a large volume of data stored in many databases, data warehouses or any other information repositories [1]. This technique is a highly interdisciplinary area spanning from a range of disciplines like Statistics, Machine Learning, Databases, Pattern Recognition and others. Different terms are being used for data mining technique in the literature, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many data mining applications contains the dataset with huge amount of data. The increase of data size in terms of number of instances and number of features become a complex task to perform the analysis. This is due to the curse of dimensionality. Dimensionality reduction is an important research area in the field of pattern recognition, machine learning, data mining and statistics. The main objective of dimensionality reduction is to transform the high dimensional data samples into the low dimensional space such that the intrinsic information contained in the data is preserved. Once the dimensionality gets reduced, it helps to improve the robustness of the classifier and it reduces the computational complexity. Figure 1 shows the dimensionality reduction process.Dimensionality reduction can be achieved in two different ways namely, Feature selection and Feature extraction.

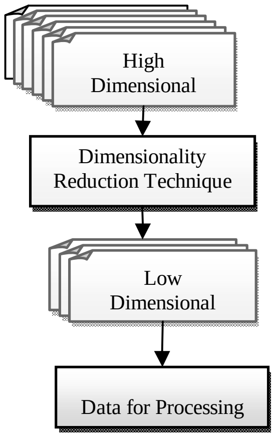


Figure 1. Dimensionality Reduction Process

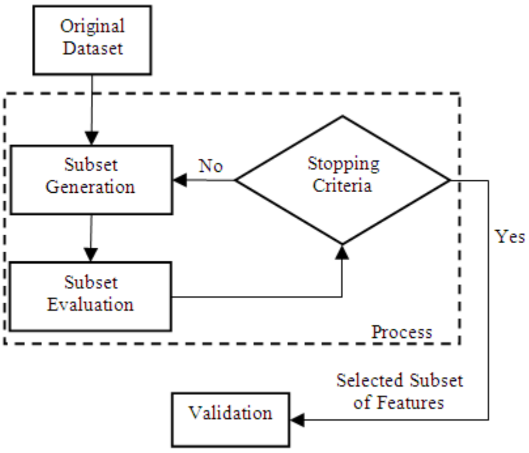
2. FEATURE SELECTION

Feature selection is a technique which is used to find the good quality of relevant features from the original dataset using some objective measures. Nowadays, Feature Selections have become challenging issues in the field of Pattern Recognition [2] Machine Learning [3], Data Mining [4] and Case-Based Reasoning [5].

Feature Selection is a process of finding an optimal or suboptimal subset of x features from the original X features. It requires a large search space to obtain the optimal feature subset. The optimal feature subset is measured by evaluation criteria. The main objective of the feature selection is to reduce the number of features and to remove the irrelevant, redundant and noisy data [6]. By reducing the features, one can reduce the system complexity, overfitting of learning methods and increase the computational speed.

3. FEATURE SELECTION PROCEDURE

The feature selection procedure includes four important key steps; subset generation, subset evaluation, stopping criterion and result validation [7] which are shown in Figure 2.



3.1 Subset Generation

It is a search process that generates the candidate feature subset using certain search strategy. The process has two basic issues. They are, search direction and search strategy. Firstly, a starting point must be selected which in turn influences the search direction. The search directions are divided into forward search, backward search and bi-directional search [Liu, 1998]. The search process starts with an empty set and adds the features progressively one by one (forward search) or starts with full sets and removes the features one by one (backward search) or starts with both ends and adds and removes the features simultaneously(bi-directional search). Secondly, a search strategy must be decided. The search strategies are categorized into complete search, sequential search and random search [7].

3.2 Subset Evaluation

In subset evaluation, the evaluation criterion is used to evaluate each newly generated subset. The evaluation criterion is used to determine the goodness of the subset (i.e., an optimal subset selected using one criterion may not be optimal according to another criterion). The evaluation criteria are divided into Independent, Dependent and Hybrid criteria[8]

3.3 Stopping Criteria

It is used to stop the feature selection process. The feature selection process may stop under one of the following criteria [7]. A predefined number of features is selected,
A predefined number of iterations is reached, In case, addition (or deletion) of a feature fails to produce a better subset, An optimal subset according to the evaluation criterion is obtained.

3.4 Validation

The validation process is used to measure the resultant subset using the prior knowledge about thedata. In some applications, the relevant features are known beforehand, a comparison is done between the known set of features with the selected features [Lad, 2011]. However, in most real-world applications, the prior knowledge about the data is not available. In such case, the validation task is performed by an indirect method. For example, the classifier error rate test is used as an indirect method to validate the selected features. The error rate on the full set of features and the same on the selected set of features are compared to find the goodness of the feature subsets.

4. FEATURE SELECTION APPROACHES

The feature selection approaches are broadly classified into three types. They are, Filter Approach [8], Wrapper Approach [9] and Hybrid Approach [10].

4.1 Filter Approach

In Filter approach [12], a statistical measure is used as a criterion for selecting the relevant features. This approach is computed easily and very efficiently.

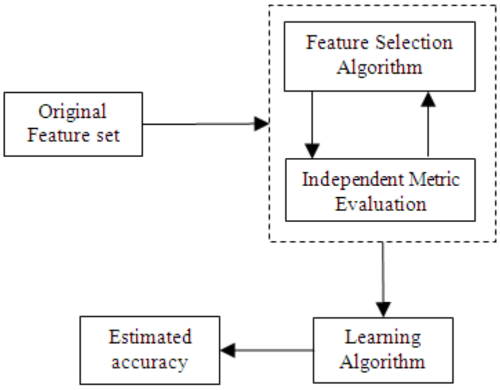


Figure 3. Filter Approach in Feature Selection

4.2 Wrapper Approach

In Wrapper approach [12] a learning algorithm is used to select the relevant features.

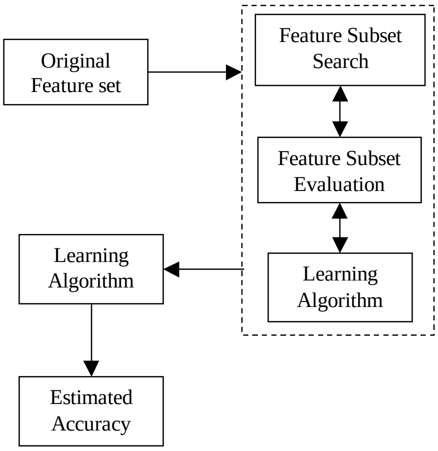


Figure 4. Wrapper Approach in Feature Selection

4.3 Hybrid Approach

The Hybrid approach is developed by combining the above filter approach and wrapper approach to handle larger datasets. In this approach the feature set is evaluated using both independent measure and a data mining algorithm. The independent measure is used to choose the best subset for a given cardinality and the data mining algorithm selects the finest subset among the best subsets across diverse cardinalities [13]. Figure 5 shows the Hybrid approach.

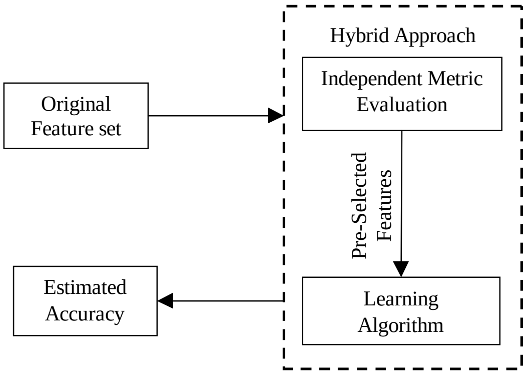


Figure 5. Hybrid Approach in Feature Selection

5. Statistical Measure used to find the Feature Selection

5.1 Information Theory

Information theory provides a platform for various machine problems. In feature selection information theory methods are applied in the filter method. The usage of information theory is found in various feature ranking measures. In this theory, the filter methods use ranking criteria based on the statistical measures which are shown below. In the following formulation F_i refer to features and c_1 and c_2 refer to classes [14].

Shannon Entropy

$$(F_i) = -\int p(f|c_1) log p(f|c_1)\, df - \int p(f|c_2) log p(f|c_2)\, df \qquad ...(1)$$

Euclidean Distance

$$E(F_i) = \sqrt{p(f|c_1 - p(f|c_1))^2 df} \qquad ...(2)$$

Kolmogorov Dependence

$$KO(F_i) = \int (p(f|c_1)) - (p(f|c_2)) df \qquad ...(3)$$

Entropy Measure

$$H(X|Y) = \sum_{y \in Y} p(y) log_2 (p(y)) \qquad ...(4)$$

5.2 Mutual Information

Mutual information provides suitable criterion for featureselection. Firstly, the mutual information is a measure used forreducing the uncertainty of the class label. Secondly,maximizing the mutual information between the features andthe class labels minimizes a lower bound on the classificationerror [4].

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right) \qquad ...(5)$$

where

p(x, y) - joint probability distribution function of X and Y.
p(x) and p(y) - marginal probability distribution function of X and Y.

5.3 Information Gain (IG)

It is a symmetrical measure. The information gained about Y after observing X is equal information gained about X after observing Y. The formulation is given below [14].

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \qquad ...(6)$$

5.4 Gain Ratio (GR)

It is a non- symmetrical measure. The IG measure is normalized by dividing by the entropy of X. By this normalization the GR measure has only two values 0 and 1. If it indicates 1, then it shows that X completely predicts on Y. If it has the value 0, there is no relation between Y and X [14]. The GR measure is defined by

$$GR = \frac{IG}{H(X)} \qquad ...(7)$$

5.5 Symmetric Uncertainty (SU)

The symmetric uncertainty measure compensate for the inherent bias of IG by dividing the entropies of X and Y. The SU falls to the range 0 and 1. SU = 0 indicates that X and Y are uncorrelated, and SU = 1 means that the knowledge of one attribute completely predicts the other [14]

$$SU = 2X \frac{IG}{H(Y) + H(X)} \qquad ...(8)$$

5.6 Correlation- Based Feature Selection

It ranks the feature based on the correlation measure. The formulation is shown below [14]

$$M_s = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \equiv \frac{\sum_K SU(X_K, C)}{\sqrt{\sum_i \sum_k SU(X_i, X_k)}} \qquad ...(9)$$

where

Ms - heuristic merit of a feature set
C - Class attribute
$\bar{r}_{cf}$ - mean feature-class correlation
$\bar{r}_{ff}$ - average feature-to-feature inter-correlation

5.7 χ^2- Statistics (CHI)

The chi-square statistics evaluates the correlation betweentwo variables and determines whether they are independent orcorrelated [4]. The test result determines whether they arepositively correlated or not

6. Feature Extraction

The feature extraction technique is used to obtain the most relevant information from the originaldata and represent that information in a lowerdimensionality space. This technique is used to select a new set of features. Thefeature transformation may be a linear or nonlinear combination of original features. The features are extracted using the following methods.

6.1 Principal Component Analysis

Principal Component Analysis (PCA) [15] is a classical statistical technique which is widely used to reduce the dimensionality of a dataset consisting of enormous amount of interrelated variables. PCA reduces the dimensionality by transforming the original dataset into a new set of variables, called principal components, where the largest variance present in the original dataset is captured by the highest component in order to extract the most important information. To show the workings of PCA, consider two dimensional dataset (p, q) with fifty observations. Figure 6 shows the plot of fifty observations on the two variables p, q that are highly correlated and Figure 7 show the transformed dataset using these principal components.

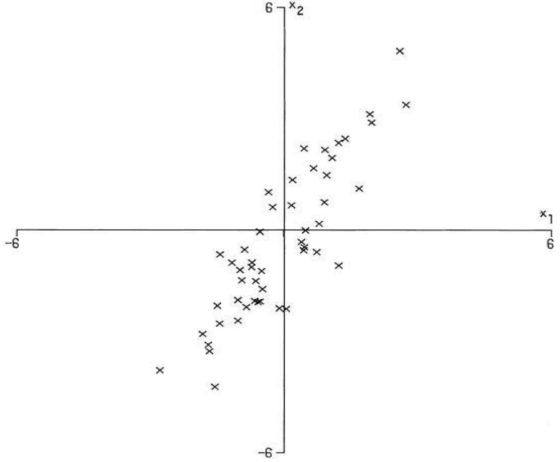


Figure 6 Plot of 50 Observations on Two Variables p, q

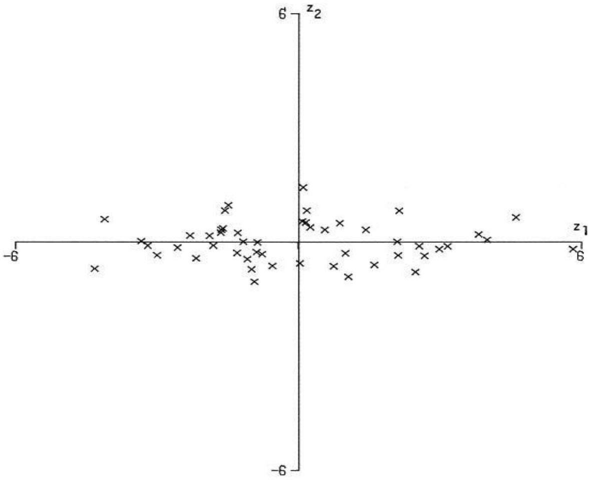


Figure 7 Transformed Dataset of Figure 6

The computational steps of the PCA algorithm is given below,

Step 1 Calculate the Mean:

$$\bar{y} = \frac{1}{M}\sum_{i=1}^{M} y_i$$

Step 2 Subtract the Mean from variables $: \Phi_i = y_i - \bar{y}$

Step 3 Form the Matrix A= [Φ1, Φ2,…. ,ΦM] (N x M matrix), then compute:

$$C = \frac{1}{M}\sum_{p=1}^{M} \Phi_p \Phi_p^T = AA^T$$

(sample covariance matrix, *NxN*, characterizes the scatter of the data)

Step 4 Calculate the Eigenvalues of $C: \lambda_1 > \lambda_2 > \cdots > \lambda_N$

Step 5 Calculate the Eigenvectors $C: u_1, u_2, \ldots, u_N$

Since C is symmetric, $u_1, u_2, \ldots, u_N$ form a basis, (i.e., any vector *x* or actually $y - \bar{y}$, can be written as a linear combination of the Eigenvectors):

$$y - \bar{y} = b_1 u_1 + b_2 u_2 + \cdots + b_N u_N = \sum_{i=1}^{N} b_i u_i$$

Step 6 (dimensionality reduction step) keep only the terms corresponding to the K largest Eigenvalues:

$$\hat{y} - \bar{y} = \sum_{i=1}^{K} b_i u_i \quad where\ K \ll N$$

6.2 Principal Feature Analysis

The Principal Feature Analysis (PFA) [16] technique is derived from the Principal Component Analysis (PCA). It is an unsupervised technique. Let P be a zero mean m-dimensional random feature vector and consider X be the covariance matrix. Let D be a matrix whose columns are the orthonormal Eigenvectors of the matrix X.

$$X = D\Lambda D^T$$

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{bmatrix}$$

where λ1, λ2,….., λm, be the eigenvalues of X. Let Dq be the first q column of D. Let V1, V2, V3…Vn Є Rm be the rows of Dm. Each vector Vi represents the projection of the ith feature of the vector P. To find the best subset, row vector Vi is used to cluster the features which have a high correlated measure. Finally, relevant features are obtained from each cluster to form a feature subset. The algorithm is summarized in the following five steps:

Step 1: Calculate the sample covariance matrix or true covariance matrix. In few cases correlation matrix is preferred to use instead of covariance matrix. The correlation matrixes is computed by

Step 2: Calculate the Principal components and eigenvalues of the Covariance/ Correlation matrix by

Step 3: Matrix Dq is constructed from D by choosing the subspace dimension q. This can be chosen by deciding how much of the variability of the data is desired to be retained.

Step 4: Cluster the vector into p clusters using the K-Means algorithm. Euclidean distance is used as a distance measure in K-Means algorithm.

Step 5: Find the corresponding vector Vi from each cluster which is closest to the mean of the cluster. Choose the corresponding features, Si, as a principal feature. This step yields the choice of p features.

6.3 Fisher Criterion

Fisher score is one of the simplest criterions used for feature selection algorithms [17]. In this criterion, the features having the similar values in the same class and the dissimilar values in different classes are selected. The Fisher score is calculated using the formula in equation (2.1)

$$FS = \frac{\sum_{k=1}^{m} s_k\left(\mu_{i,k} - \mu_i\right)^2}{\sum_{k=1}^{m} s_k \sigma_{i,k}^2}$$

where,
μ_i is the mean of the features,
s_k is the number of samples in the kth class,
$\mu_{i,k}$ is the mean of the features in the kth class,
$\sigma_{i,k}^2$ is the variance of the features in the kth class.

6.4 Linear Discriminant Analysis (LDA)

In many domains like Data Mining, Machine Learning, and Pattern Recognition, high dimensional data are commonly available. Acquiring valuable knowledge in high dimensional spaces is a challenging task. In this scenario, the data points are far apart from each other and the similarities between data points are difficult to compare and analyze [14]. LDA is an important dimensionality reduction method to handle the high dimensional data. This technique mainly projects the high-dimensional data into lower dimensional space. LDA aims to maximize the between-class distance and minimize the within-class distance in the dimensionality-reduced space. The LDA is computed by the following equation (2.2).

$$f(X) = trace\left((X^T S_w X)^{-1}(X^T S_b X)\right) \qquad \ldots(2.2)$$

where
S_b is the between-class matrix
S_w is the within-class matrix,

$$S_b = \frac{1}{n}\sum_{i=1}^{m} k_i\,(c_i - c)(c_i - c)^T$$

$$S_w = \frac{1}{n}\sum_{i=1}^{m}\sum_{x \in X_i}(x - c_i)(x - c_i)^T$$

where,
X_i is the index set of ith class,
c_i is the mean vector of ith class.

7. Conclusion

In this paper, a study has been carried out to know how the high dimensionality problem has been solved using different dimensionality reduction Techniques. This paper gives a complete knowledge about how the features was selected and extracted using feature selection and feature extraction techniques respectively. In feature selection technique the most relevant features are selected using statistical measure and some of the statistical measure are explained in detail. In feature extraction technique the new feature were obtained from the original features using the various statistical techniques and most popular statistical techniques were explained in detail. Hence this paper will help the beginners who were doing research in the dimensionality reduction techniques.

References

[1] AshaGowdaKaregowda, M.A.Jayaram, A.S. Manjunath,“Feature Subset Selection Problem using Wrapper Approach in Supervised Learning”, International Journal of Computer Applications, Volume 1,No. 7, 2010, pp.13-17, ISSN:0975 – 8887.

[2] Sven F.Crone, NikolaosKourentzes, “Feature selection for time series prediction–A combined filter and wrapper approach for neural networks”, Journal of Neurocomputing, Volume 73, Issues 10-12, June-2010,pp. 1923-1936, ISSN: 0925-2312.

[3] Pawel Smialowski1,Dmitrij Frishman1,and Stefan Kramer, "Pitfalls of supervised feature selection", Published by Oxford University Press, December 9 , 2010.

[4] Subramanian Appavu Alias Balamurugan, RamasamyRajaram,

[5] “Effective and Efficient Feature Selection for Large-scale Data Using Bayes' Theorem”, International Journal of Automation and Computing, Volume6, Issue 1, Feb 2009, pp. 62-71.

[6] Yuanhong Li, Ming Dong, and Jing Hua, “A Gaussian Mixture Model To Detect Clusters Embedded In Feature Subspace”, Journal of Communications in Information and Systems, Volume 7, Number. 4, 2007, pp. 337-352.

[7] Peng Liu, Naijun Wu, Jiaxian Zhu, Junjie Yin, and Wei Zhang, “A Unified Strategy of Feature Selection”,The Second International Conference on Advanced Data Mining and Applications(ADML 2006), China, August 2006, pp. 457 – 464.

[8] Frederico Coelho, Antonio Padua Braga, and Michel Verleysen, “Multi-Objective Semi-Supervised Feature Selection and Model Selection Based on Pearson’s Correlation Coefficient”, Springer LNCS 6419, 2010, pp. 509–516.

[9] IanisseQuinzán, José M. Sotoca, FilibertoPla, “Clustering-based Feature Selection in Semi-supervised Problems”, Ninth International

Conference on Intelligent Systems Design and Applications, Italy, 2009, pp. 535- 540, ISBN: 978-0-7695-3872-3/09

[10] Jidong Zhao, Ke Lu, Xiaofei He, "Locality sensitive semi-supervised feature selection", Journal of Neurocomputing, Volume 71, Issues 10-12, June-2008, pp. 1842-1849, ISSN: 0925-2312. [10]

[11] M. Ramaswami and R. Bhaskaran, "A Study on Feature Selection

[12] Techniques in Educational Data Mining", Journal of Computing Volume1, Issue 1, December 2009, pp.7-11,ISSN: 2151-9617.

[13] Zhu Zhang,"Mining relational data from text: From strictly supervised toweakly supervised learning", Journal of Information System, Volume33, issues 3, May 2008, pp 300-314, doi:10.1016/j.is.2007.10.002

[14] HuanLiu,HiroshiMotoda, Rudy Setiono and Zheng Zhao, "FeatureSelection: An Ever Evolving Frontier in Data Mining",Journal ofMachine Learning Research, volume 10, june 2010, Hyderabad, pp. 4-13.

[15] Le Song, Alex Smola, Karsten M. Borgwardt, Justin Bedo, "SupervisedFeature Selection via Dependence Estimation", procedings Internationalconference of Machine Learning(ICML), June 2007, USA.

[16] Seoung Bum Kim, PanayaRattakorn, “Unsupervised feature selectionusing weighted principal components”, International journal of ExpertSystems with Applications, Volume 38 Issue 5, May, 2011, pp. 5704- 5710.

[17] DaoqiangZhanga,Songcan Chena, Zhi-Hua Zhoub, “ Constraint Score:Anew filter method for feature selection with pairwise constraints” ,Elsevier: The Journal of the Pattern Recognition Society, October 2007,DOI: 10.1016/j.patcog.2007.10.009.

AUTHOR

Dr. V. Arul Kumar is working as Assistant Professor in the Department of Computer Science, Thanthai Hans Roever College, Perambalur, Tamil Nadu, India. He has 2 years of experience in teaching and 4 years of experience in research. He has published 21 research articles in the International / National Conferences and Journals. His research interests are: Data Mining, Cryptography and Cognitive Aspects in Programming.

N. Elavarasan is working as Assistant Professor in the Department of Computer Applications, Thanthai Hans Roever College, Perambalur, Tamil Nadu, India. He has 14 years of experience in teaching. He hasauthored books on “Web Metrics” and “Advanced Visual Basics 6.0”. He is currentlypursuing doctor of philosophy programme at Nehru Memorial College (Autonomous), Puthanampattiand his current area of research isData mining.