

FEATURE SELECTION FOR SHARE MARKET DATASET USINGG ENTROPY AND INFORMATION GAIN

Aruna M, Dr. Sugumaran S, Dr. Srinivasan V
Associate Professor, Assistant Professor, Professor
Department of Computer Applications
Velalar Colleg of Engineering and Technology,
Erode, Tamilnadu-638009, India.

Abstract: The research works that are being carried out recently undergoes a very huge dataset due to the availability of internet and vast storage area. Every dataset has millions of records with large attributes, which becomes difficult for each researcher to down size the dataset so that they could give the result of their finding accurately. In this paper using the entropy and information gain concept the large dataset is reduced to the researcher so that one could find the result accurately with the minimal attributes. Share dataset is used to show how the results are varied by choosing the attributes manually. So to attain the accurate result the concept of entropy and the information gain is used for the share market dataset to predict the result accurately.

Keywords: Large database, Entropy, Information Gain, Feature Selection, Reduced Dataset, Attribute Selection etc.

I. INTRODUCTION

Feature selection's main idea is to select a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models [1]. It is often the case that finding the correct subset of predictive features is an important problem for clustering accuracy. Many factors affect the success of data mining algorithms and the quality of data is one such factor. The discovery during training becomes difficult if information is irrelevant or redundant [2], or the data is noisy and unreliable. Attribute selection identifies and removes the redundant information as possible. Learning algorithms differ in the amount of emphasis they place on attribute selection [3]. The size of the hypothesis space is reduced by reducing the dimensionality of the data and allows algorithms to operate faster and more effectively. Accuracy on future clustering can also be improved in some cases.

A feature selection algorithm can be viewed as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. Feature extraction and feature selection are two approaches to dimension reduction.

1.1 Feature extraction

Feature extraction actually transforms the attributes. The original attributes are linear combinations of the transformed attributes or features. The feature extraction process results in a much smaller and richer set of attributes [4]. The maximum number of features is controlled by setting for feature extraction models. Models built on extracted features may be of higher quality, because the data is described by fewer with more meaningful attributes [13]. A data set with higher dimensionality can be projected onto a smaller number of dimensions with feature extraction. Since a complex data set can be effectively visualized when it is reduced to two or three dimensions which is useful for data visualization,. Some applications of feature extraction are latent semantic analysis, data compression, data decomposition, projection and pattern recognition. The speed and effectiveness of supervised learning can be enhanced with Feature extraction

1.2 Feature selection

Feature selection works by calculating a score for each attribute, and then selecting only the attributes that have the best scores [5]. One can also adjust the threshold for the top scores. Analysis services provide multiple methods for calculating these scores, and the exact method that is applied in any model depends on these factors:

- The algorithm used in the model
- The data type of the attribute
- Any parameters that may be set on the model
-

The Feature selection is applied to predictable attributes, inputs, or to states in a column. When scoring for feature selection is complete, only the attributes and states that the algorithm selects are included in the model-building process and can be used for prediction [6]. Attribute selection techniques can be categorized according to a number of criteria [12]. Some popular categorizations are filter, wrapper and embedded methods to describe the nature of the metric used to evaluate the worth of attributes. The feature selection becomes more important for any algorithm to present their finding accurately [9]. The result various with the selection of attributes, in this paper entropy and information gain is used to select the meaningful attributes and finally use the clustering concept to show that the attribute selection gives different results on selection of different attributes.

1.3 Materials and Methods

Feature selection implies not solely cardinality reduction, which means imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, but also the selection of attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis. The ability to use feature selection is vital for effective analysis, because datasets frequently contain far more information than is needed to build the model. For example, a dataset might contain 100 columns that describe the characteristics of customers, but if the data in some of the columns is very sparse that would gain very little help from adding them to the model. If the analysis keeps the unneeded columns while building the model, more CPU and memory are required during the training process, and more storage space is required for the completed model. One should typically need to remove unneeded columns because they might degrade the quality of discovered patterns, for the following reasons:

- Some columns are noisy or redundant which makes it more difficult to discover meaningful patterns from the data.
- To discover quality patterns, many data mining algorithms require much larger and high-dimensional training data set. But the training data is not sufficient in some data mining applications.

This paper uses Shannon's entropy for selection of attributes, the share market dataset has more than 25 attributes from which few attribute is select for faster clustering and to produce accuracy result.

1.4 Shannon's Entropy

In this paper Shannon's Entropy is used which is a fastest attribute ranking methods and is often used in clustering techniques. The proposed system need to select the attribute which gives more information for clustering from a large number of attributes, to obtain the accurate clustering. The gain Entropy is obtained from Information Theory [10]. Equations 1 and 2 are used to calculate the entropy and the information gain. As the entropy and the information gain is calculated, one selects the threshold point for the selection of the attributes, if the dataset contains more than five attributes and this are order from the highest values based on entropy. The attribute that contains the highest values are considered for the clustering process by leaving other attributes which does not give more information for the evaluation process [11]. In addition to entropy and the information gain a new method is used to select the attributes, if the attribute contains same value for the ranking, only one attribute is selected for the attributes which have the same value. The section below shows how the attribute selection for share market dataset is done using entropy and the information gain for the evaluation process.

1.5 Entropy

Entropy is a measure of variability in a random variable. It will measure how the particular attribute divides the training examples into the number of result classes [20]. Table 1 show the sample dataset for the share market data [21]. In share market dataset one need to select the attribute which gives more information for clustering so as to make accurate clustering. For defining gain Entropy is obtained from Information Theory. Entropy is used to calculate the amount of useful information in an attribute. This is calculated as Equation. (1)

$$\text{Entropy (S)} = -\sum P(x_i) \log_b P(x_i) \quad (1)$$

Where:

S	=	Collection of Samples
x_i	=	Set of outcomes
$P(x_i)$	=	Proportion of S to the class x_i

Entropy was used by J. Ross Quinlan who has used Entropy in ID3 algorithm. This algorithm is based on the Concept Learning System (CLS).

1.6 Information Gain (IG)

The information gain relies on the decrease in entropy once a dataset is split on an attribute. First the attribute that creates the most homogeneous branches are identified Equation (2)

$$\text{IG}(Y/X) = H(Y) - H(Y/X) \quad (2)$$

The entropy calculated for the dataset of share market dataset are ordered as public share holding with 0.97, high-average 0.92, 52 weeks high 0.88, Average of 52 weeks 0.86 and so on.

Table 1.1. Sample database of partial record is shown from the Indian share market dataset

S.No	A1	A2	A3		A13	A14
1	Ashok Leyland	1.00	1.00		1.00	0.98
2	Asian paints	0.04	0.05		0.04	0.68
3	Bank of Baroda	0.06	0.05		0.07	0.84
	...		...			
542	Titan Inds	0.11	0.12		0.11	0.95
543	Torrent Power	0.04	0.03		0.03	0.55
544	UltratechCem	0.04	0.03		0.04	0.48

A1-Share name, A2-Q1-Profit from April to June, A3-Yearly profit, A4-52 weeks high, A5-Average of q1 to q4, A6-Average profit of the year, A7-Average of 52 weeks, A8-Q2- profit from July to September, A9-Q3-Profit from October to December, A10-Q4-Profit from January to March, A11-Public share holding, A12-52 weeks low, A13-High-Average, A14-Average total.

The selected attributes alone is taken for further process in clustering the share market dataset into three categories of high, Medium and low.

II. CLUSTERING METHODS

Clustering is the method of grouping a set of data objects into multiple teams or clusters so that objects inside a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute value, users describing the objects and often involve distance measures [22]. Clustering as a data mining tool has its roots in many application areas such as biology, security, business intelligence, share market and web search etc.

From a machine learning perception clusters correspond to unknown patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an exceptional role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, share market and many others. Clustering is the subject of dynamic research in several fields such as statistics, pattern recognition, and machine learning. Data mining adds to clustering the complications of very large datasets with many attributes of different types [23]. This imposes unique computational requirements on relevant clustering algorithms. There are different clustering methods evolved they are partitioning methods, hierarchical methods, density based methods and grid based methods.

2.1 Partitioning methods

Partitioning is the simplest and most fundamental version of cluster analysis, which organizes the objects of a set into several exclusive groups or clusters. To keep the problem specification concise, one can assume that the number of clusters is given as background knowledge which is the starting point for partitioning methods. Formally, given a data set, D , of n objects, and k clusters, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions ($k \leq n$), where each partition represents a cluster[24]. The clusters are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are similar to one another and dissimilar to objects in other clusters in terms of the data set attributes. The commonly used partitioning methods are K-means and K-medoids.

2.2 Hierarchical Methods

A hierarchical clustering method works by grouping data objects into a hierarchy or tree of clusters. Representing data objects in the form of a hierarchy is useful for data summarization and visualization. This builds a cluster hierarchy or, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring 6 data on different levels of granularity [25]. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and Divisive (top-down). An agglomerative clustering starts with one point (singleton)

clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion is achieved.

2.3 Density-Based Methods:

To find the clusters of arbitrary shape such as the “S” shaper and oval clusters. They would likely inaccurately identify convex regions, where noise or outliers are included in the clusters. An open set in the Euclidean space can be divided into a set of its connected components. The implementation of this idea for partitioning of a finite set of points requires concepts of density, connectivity and boundary. They are closely related to a point is nearest neighbors. A cluster, defined as a connected dense component, grows in any direction that density leads. Therefore, density-based algorithms are capable of discovering clusters of arbitrary shapes [26]. Also this provides a natural protection against outliers. The basic techniques of density based clustering methods are DBSCAN, OPTICS and DENCLUE.

2.4 Grid-Based Methods:

A grid based clustering method takes a space driven approach by partitioning the embedding space into cells independent of the distribution of the input objects. The grid based clustering approach uses a multi resolution grid data structure. This quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is its fast processing time, which is typically independent of the number of data object [27]. The methods used for density based clustering are STING and CLIQUE.

In this paper the partitioning method is used for the share market dataset to cluster as high, medium and low.

III. IMPLEMENTATION

Share market dataset was framed from getting the information from the National Stock Exchange (NSE) is India's leading stock exchange covering various cities and towns across the country. NSE was set up by leading institutions to provide a modern, fully automated screen-based trading system with national reach [7]. The Exchange has brought about unparalleled transparency, speed & efficiency, safety and market integrity. This has set up facilities that serve as a model for the securities industry in terms of systems, practices and procedures [8].

This dataset was taken to show whether a particular share will raise or go down in future coming year based on the previous year recorded information [14]. The different attributes for the dataset are Q1 profit of April to June, Q2 profit of July to September, Q3 profit of October to December, Q4 profit of January to March, Yearly profit, Public share holding, 52 Weeks Low, 52 Weeks High, Average Profit of Q1 to Q4, Average profit of the Year, Average of 52 weeks and so on. To reduce the attribute, Entropy and the information gain are calculated for each attribute with the equation (1) and (2). These attributes are then ordered according to the highest entropy and the information gain [15] [16]. Finally only select attributes are taken for the final clustering process from a complex dataset.

In this paper the share market dataset which has 16 attribute is reduced to four attributes which gives meaningful information for clustering to be classified as high, medium and low leaving other attributes that does not give more information for the clustering purpose. The selected attributes are listed in the table below this is done based on the entropy and information gain [17].

Table 3.1. Sample Share Market Dataset

S. No	A6	A2	A7	A8	Total
1	26.294	0.497	0.351	0.352	27.494
2	34.005	0.013	0.033	0.043	34.094
3	8.947	11.606	13.298	11.330	45.181
4	5.164	7.866	6.033	8.051	27.114
	...		...		...
281	12.278	0.075	0.065	0.037	12.455
282	22.545	0.404	0.365	0.465	23.779
283	19.674	1.560	1.039	0.885	23.158
284	12.232	0.133	0.447	0.320	13.132
	...		...		...
541	22.985	1.241	1.201	0.955	26.382
542	13.527	0.836	0.898	0.989	16.250
543	11.233	0.006	0.003	0.004	11.246
544	18.212	0.598	0.631	0.698	20.139

A6 – Public share holding
 A2 – Yearly profit
 A7 – Average profit of the year
 A8 – 52_weeks high

The above share market dataset is clustered as high, medium and low based on the weight and selecting the center point based on the mean of the weight. Below diagram shows the three different clustering of share market dataset to predict high, medium and low.

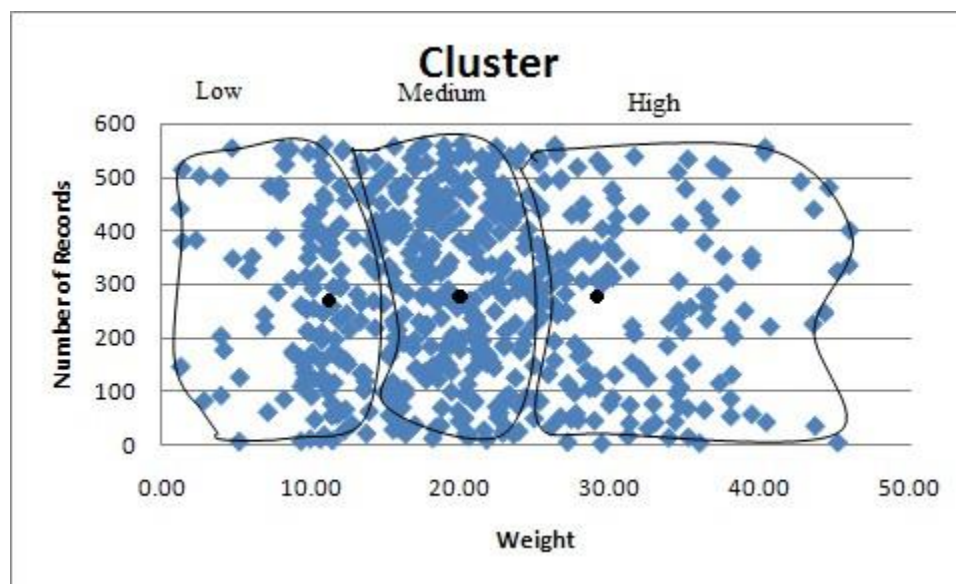


Figure 3.1: Shows the clustering of Share Market Dataset.

The above diagram shows that if the customers invest the money in the high cluster of share then they could gain more profit in the future coming years. The above result varies if the selection of attribute is different by manual selection and they would end up with less profit. That is why the entropy method is used to get more information about the attribute by selecting these attributes with more information helps to cluster the share data set with better accuracy [18] [19].

IV. CONCLUSION

This paper shows that the selection of attribute place an important role in any of the research activities, if the attribute selection is correct then the result finding will have more accuracy to be produced for any of the chosen dataset. For the share market dataset four attributes with highest information gain are chosen. They are A6 – Public share holding, A2 – Yearly profit, A7 – Average profit of the year and A8 – 52_weeks high. Even though other attributes give the values but do not have more information than these four attributes. So it is important to select these attributes that give more information from a large number of attributes. Using this entropy method four attributes are selected for the final process of clustering, with these attributes weight is calculated and based on the weight center point for each cluster is made and finally the clusters are separated as high, medium and low.

V. REFERENCES

- [1] Biesiada, J & Duch, W 2008, 'Feature selection for high-dimensional data: A Pearson redundancy based filter', *Advanced Soft Computing*, vol. 45, pp. 242-249.
- [2] Dingsheng, W, Xiang, R, & Yuting, H 2010, 'Data mining algorithmic research and application based on information entropy', *Pattern Recognition*, vol. 43, no. 1, pp. 5-13.
- [3] Gheyas, I & Smith, L 2010, 'Feature subset selection in large dimensionality domains', *Pattern Recognition*, vol. 43, no. 1, pp. 5-13.
- [4] Vandan Kuzhali, J & Srinivasan, V, 'Feature Selection Algorithm using Fuzzy Rough Sets for Predicting Cervical Cancer Risks', *International Journal Modern Applied Science* Vol. 4, Issue 9, 10 sep 2010.
- [5] Guyon, I & Elisseeff, A 2003, 'An introduction to variable and feature selection', *Journal of Machine Learning Research*, vol. 3, no. 7, pp. 1157-1182.
- [6] Halperin, E & Karp, RM 2005, 'The minimum-entropy set cover problem', *Theoretical Computer Science*, vol. 348, no. 2, pp. 240-250.
- [7] Indian Stock Exchange 2003, Stock Board information. Available from <<http://www.stocksabroad.com>> [20 June 2003].
- [8] Indian Stock Exchange 2005, Share Market Information. Available from <<http://www.nse-india.com>> [10 May 2005].

- [9] Srinivasan, V & Aruna, M, 'Classify the student with missing value to calculate future semester result for placement record using knowledge acquisition' National Journal Vol. 3 Issue No.3 2010.
- [10] Jiawei, H & Micheline, K 2006, 'Data Mining: Concepts and Techniques, 2nd edition', Morgan Kaufmann Publishers.
- [11] Kedarnath, JB. & Nur, AT 2007, 'Relationship between entropy and test data compression', IEEE Transaction on Computer Aided Design of Integrated Circuits and Systems, vol. 26, no. 2, pp. 386-395.
- [12] Srinivasan, V & Aruna, M, 'A fuzzy fast classification for share market database with lower and upper bounds' American journal of Applied Science, vol.12, sep 2012, PP.1934-1939.
- [13] Liu, H 2005, 'Evolving feature selection', IEEE Intelligence System, vol. 20, no. 6, pp. 64-76.
- [14] Neelima, B, Jha, CK & Sandeep KB 2012, 'Application of Neural Network in Analysis of Stock Market Prediction', International Journal of Computer Science & Engineering Technology, vol. 3, no. 4, pp. 61-68.
- [15] Richard, J & Qiang, S 2007, 'Fuzzy-Rough Sets Assisted Attribute Selection', IEEE Transactions on Fuzzy Systems, vol. 15, no. 1, pp. 73-89.
- [16] Srinivasan, V & Aruna, M, 'Fuzzy Classification to Classify the Income Category Based on Entropy', Journal of Computer science and Technology Vol. 11 No.2 October 2011.
- [17] Slezak, D 2002, 'Approximate Entropy Reducts', Fundamenta Informaticae, vol. 53, no. 3, pp. 365-390.
- [18] Turiel, A & Vicente, CJP 2003, 'Multifractal geometry in stock market time series', Physica A: Statistical Mechanics and its Applications, vol. 322, no. 1, pp. 629-649.
- [19] Srinivasan, V & Aruna, M, 'Fuzzy Fast Classification Algorithm with Hybrid of ID3 and SVM' International journal of Intelligence and Fuzzy System, Vol.24, May 2013, pp.556-561.
- [20] Yucel, S, Arnold R & YunTong, W 2004, 'Value of Information Gained From Data Mining in the Context of Information Sharing', IEEE Engineering Management, vol. 51, no. 4, pp. 441-450.
- [21] Zabir, HK, Tasnim, SA & Md, AH 2011, 'Price Prediction of Share Market using Artificial Neural Network', International Journal of Computer Applications, vol. 22, no. 2, pp. 0975-8887.
- [22] Srinivasan, V & Aruna, M, 'A Fuzzy Approach to Replace the Missing Data in Large Dataset', International Journal of Applied Engineering and Research, Vol.10, N0.38, May 2015, pp.28312-28317.
- [23] M. Setnes, "Supervised Fuzzy Clustering for Rule Extraction," IEEE Trans. Fuzzy Systems, vol. 8, pp. 416-424, 2000.
- [24] R.J. Hathaway, J.C. Bezdek, and Y. Hu, "Generalized Fuzzy C-Means Clustering Strategies Using Lp Norm Distances," IEEE Trans. Fuzzy Systems, vol. 8, no. 5, pp. 576-582, 2000.
- [25] K.L. Wuand M.S.Yang, "Alternative C-Means Clustering Algorithms," Pattern Recognition, vol. 35, pp. 2267-2278, 2002.
- [26] J. Seo and B. Shneiderman, "Interactively Exploring Hierarchical Clustering Results," Computer, vol. 35, no. 7, pp. 80-86, July 2002.
- [27] R. Nock and F. Nielsen, "On Weighting Clustering," IEEE Trans.Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1223-1235, Aug. 2006.