Cotton Fiber Quality Analysis System Using Fuzzy C-Means Clustering Algorithm

E. Iswarya, Dr. K. Manivannan, Dr. D. Shanthi

1 PG Student, Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul.  
2 Assistant Professor Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul.  
3 Professor & Head Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul.

Received 28 January 2017; Accepted 22 May 2017; Available online 28 May 2017

Address For Correspondence:  
E.Iswarya, PG Student, Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul  
E-mail: ishuelango1412@gmail.com

ABSTRACT

Accurate specification of cotton fiber quality is an enormous support for spinning mills to buy cotton according to minimum requirements. Quality of cotton fiber is judged on many factors, such as staple length, maturity, fineness, cleanliness, stickiness, and strength, to mention. The primary objective of the system is to achieve the standardization at different levels in the cotton supply chain management in the distributed industrial networks environment with precise measurement and classification techniques. The existing system is implemented in data mining techniques to predict the quality of cotton fiber and to found the highest correlation between short fiber index and yellowness. The proposed system includes clustering method to classify the cotton quality fiber. The proposed classification techniques are such as Principal component analysis (PCA), and Fuzzy c-means clustering in big data. These proposed algorithms are implemented within the Map-Reduce based Hadoop platform to predict moisture content, micronaire, elongation and bundle strength.

KEYWORDS: cotton quality; Principal Component Analysis; Fuzzy C-means clustering

INTRODUCTION

In many countries cotton fiber is one of the most significant fiber producing plant. Cotton crop provides fiber to the textile industry. The quality of the cotton depends on different factors like fitness, stickiness, strength, maturity, cleanliness, stable length etc. A big diversity of cotton varieties of cotton fiber properties makes that a precise and complex assessment of basic fiber parameters. Moreover cotton fiber quality classification is a very important issue for traders and spinners as well as for agriculture people, who should tend to improve the cotton yield features.

Due to increasing the volume of data, big data plays an important role in all fields. There are various methodologies are developed to classify the quality of the cotton fiber. But the classification of cotton fiber quality for a distributed environment has not been proposed yet.

The cotton quality classification for distributed environment is proposed. The proposed system includes clustering methods to classify the cotton quality fiber. Then an efficient classification technique is used. This proposed method is implemented within the Map-Reduce based Hadoop platform.

Literature Survey:

Kelly, Joliffe et al [10,3] described cotton breeders which utilized an effective, sole and primary source of fiber quality data is High Volume Instrument (HVI) during the selection of plants. In addition to that Advanced...
Fiber Information System (AFIS) is used to provide additional information about cotton fiber like fiber maturity and length characteristics. However, these methods have low spinning performance and yarn quality.

Kannan et al [8] examined the gene effects to find out the fiber quality in upland cotton and potential for improvement of yield components. These are examined for twelve characters in three crosses of Gossypium hirsutum by six generation mean model. The additional characteristics like seed index, fineness, boll weight, bundle strength, lint index, uniformity ratio and ginning outturn were under the control of dominance gene action. It improves the determination of fiber quality in upland.

Kazama, Grubbs et al [9,5] analyzed the variables of cotton fiber characteristics based on the stripper, picker and manual harvest system and two cultivars utilizing statistical methods. For this analysis cotton samples were taken from storer basket of harvesters. The analyzed the fiber characteristics according to the HIV system. The principal hierarchical and non hierarchical components were grouped together into three clusters based on the cotton fiber characteristics. However this method has high initial cost and limited amount of storage.

May, Schleth et al [13,20] determine that selection for fiber strength by HVI and stelometer measurement in similar strength gains, though that interrelated responses of other properties which affect textile performance can differ. By using HIV measurement fiber strength is determined which is more similar to the determination of fiber strength using stelometer. However, the instruments are affected by moisture content and cannot detect thick and thin place.

Sachar, Norum, et al [19, 15] introduced huge, saturation and intensity (HSI) and YCbCr model for detection of cotton contaminants and cotton classification. In order to detect the different types of contaminants automatic thresholding is employed. Then the naive bayes classifier is utilized to differentiate the different types of contaminants from the fiber. The classifier used shape descriptors like area, extent, orientation and solidity as features to classify the contaminants. However, this method has high complexity and less accuracy.

Iswarya, Kruskal et al [6, 11] determine the precise detail of cotton fiber quality is a colossal support for turning factories to purchase cotton as indicated by least necessities. Cotton tests from three cotton developing districts of Uganda were described for quality parameters utilizing the High Volume Instrument (HVI). Cotton yellowness and short fiber file are found to represent a significant part of the inconstancy in cotton fiber quality. In light of Cotton Outlook's reality grouping and USDA Standards, the cotton under test was esteemed of high and uniform quality, falling amongst middling and good middling evaluations.

Abdi, Carol [1,2] described Singular Value Decomposition (SVD), which is a generalization of the Eigen value decomposition. It was used to analyze rectangular matrices. A matrix was decomposed into two matrices by analogy with the Eigen decomposition. The main intend of SVD is to decompose a rectangular matrix into one diagonal matrix and two orthogonal matrices. However, the main disadvantage of SVD is high computation cost and it is not fast from the computational point of view .

Zhang, Saporta et al [21,18] proposed piecewise linear transform model to detect the foreign fibers in cotton. The best tree analysis structure was computed contrasing to the initial one. The redundant information was deleted by optyimizing the entropy value. Thus the proposed method improves the detection of foreign fibers in cotton. However, it has high time comlexity and less efficent for detection of foregin fibers in cotton.

Pai, O’Connor et al [17,16] proposed a method for detection of cotton contaminant. The cotton contaminant may leads to wrong assessment of cotton quality. The proposed X- ray microtomographic image analysis was processed in a high degree of success for non invasive computation of cotton for recognition of contaminants in cotton. However this method consumes more time.

Militky, MacKay et al [14, 12] defined the procedure for computation of cotton quality index. It can be evaluated by simply changing for other selected properties of cotton or other set of weights. These are more significant for future cotton variables. But the major drawback of this type of evaluation is high complexity.

Kamalha, Deza et al [7, 4] introduced different machine learning methods for accurate specification of cotton fiber quality. Initially the features of the data are selected by using Principal Component Analysis (PCA) then the selected are used in Agglomerative Hierarchical clustering (AHC) and K means clustering to cluster the cotton fiber data. However this method has high time complexity.

Proposed Methodology:

In this section the proposed methodology is described in detail. The proposed methodologies are used in a distributed environment to classify the quality of the cotton fiber by using Principal Component Analysis (PCA) and Fuzzy C-means clustering. Moreover the proposed methodologies is implemented within Map Reduce based Hadoop framework. The proposed system has different advantages like complexity is less for a large amount of data, less time complexity and has high accuracy rate of classification.

Technologies currently used for accurate specification of cotton fiber quality have some limitations. These limitations may leads to wrong assessment of cotton quality fiber. Therefore the quality of cotton fiber is classified in distributed platform. The proposed system used Principal Component Analysis (PCA) for feature
extraction. Then the proposed system used Fuzzy C means clustering to classify the cotton fiber data. This proposed method is implemented within the MapReduce based Hadoop platform. The experimental results are conducted to prove the effectiveness of the proposed methodologies.

A. PCA feature extraction:
The features of the cotton fiber data is extracted based on Principal component analysis. PCA analyzed data where observations are described by several inter correlated quantitative dependent variables. It extracts more important features and presents those features as a set of new orthogonal variables named as principal components with relationships between variables and observations. PCA is described by the Eigen decomposition of positive semi-definite matrices and upon the singular value decomposition (SVD) of rectangular matrices. The overall flow of PCA is explained in the following algorithm.

Algorithm
Input: Data sets
Output: Features
1. The I x L data matrix X has a singular value decomposition
   \[ X = P \Delta Q^T \]
   where P is the Left singular vectors, Q is the Right singular vectors and \( \Delta \) is the Diagonal matrix of singular values
2. Compute the inertia of a column
   \[ Y_j = \sum_i X_{ij}^2 \]
3. Compute Euclidean distance
   \[ d_{ig}^2 = \sum_i X_{ij}^2 \]
   where \( g \) is the Centre of gravity of rows
4. Obtains I x L matrix of factor scores F
   \[ F = P \Delta \]
5. Combine step 1 and 4 to obtain projections
   \[ F = P \Delta = P \Delta Q^T Q = XQ \]
6. Calculate the matrix X
   \[ X = FQ^TX \]
7. Calculate the contribution
   \[ c_{tr}i = \frac{f_{i1}^2}{\sum_i f_{i1}^2} = \frac{f_{i1}^2}{\lambda_{i1}} \]
   where \( \lambda_{i1} \) is the Eigen value of the Ith component
8. Calculate the squared cosine
   \[ c_{os}ij = \frac{f_{i1}^2}{\sum_i f_{i1}^2} = \frac{f_{i1}^2}{d_{ig}^2} \]
9. Factorial axes are obtained

B. Fuzzy C means Clustering in distributed environment:
The basic Fuzzy c-means clustering are based on the optimization of the basic c-means objective function or some modifications of the objective function. The basic algorithm of FCM is given as follows.

Algorithm
Input: Features
Output: Clusters
Let \( X = \{x_1, x_2, x_3 \ldots, x_n\} \) be the set of data points and \( V = \{v_1, v_2, v_3 \ldots, v_c\} \) be the set of centers.
1. Randomly select cluster centers \( c \)
2. Calculate the fuzzy membership \( \mu_{ij} \)
   \[ \mu_{ij} = 1/\sum_k (d_{ij}/d_{ik})^{(2/m-1)} \]
3. Compute the fuzzy centers 'v' using:
\[ v_j = \left( \sum_{i=1}^{n} (\mu_{ij})^m x_i \right) \left( \sum_{i=1}^{n} (\mu_{ij})^m \right)^{-1}; \quad \forall j = 1, 2, \ldots, c \]

4. Repeat step 2 and 3 until the minimum J value is achieved.

In the above algorithm, \( k \) represents iteration step, \( \beta \) represents termination criteria between 0 to 1, \( U = (\mu_{ij})_{nc} \) represents the fuzzy membership matrix and \( J \) represents the objective function.

In order to enable big data processing, MapReduce based FCM is proposed. The mapping of FCM, FCM is partitioned into two MapReduce jobs. The first MapReduce job computes the centroids matrix by iterating over the cotton fiber data records. Then the second MapReduce job is required since the following computations need the complete centroids matrix as the input. In the second MapReduce process, update the membership matrix by computing the distances and it also used to calculate the fitness value. It is explained in the following algorithm.

Algorithm
Input: dataset
Output: cluster
1. Randomly initialize membership matrix
2. while stopping condition is not met do
3. vertically merge data set with membership matrix and store in HDFS
4. Hadoop calls map and reduce jobs of first and second mapreduce operation
5. copy membership matrix from HDFS and store locally
6. End while

MapReduce job 1
Input: key: data record, value: data record values and membership matrix
Output: <key',value'> pair, where value' is the intermediate centroid matrix
1. for each key do
2. calculate intermediate centroid matrix using Equation 5 and store in value'
3. emit <key',value'> pair
4. end for

MapReduce job 2
Input: key: data record, value: data record values and centroid matrix
Output: <key',value'> pair, where value' is the intermediate membership matrix
1. for each key do
2. calculate distances using Equation 7
3. update intermediate membership matrix using Equation 6 and store in value'
4. emit <key',value'> pair
5. end for
6. for each key do
7. merge intermediate membership matrices and store in value'
8. emit <key',value'> pair
9. end for

RESULT AND DISCUSSION

The performance of the proposed system is compared to the previous techniques in terms of accuracy and time complexity. For the experimental purpose, the cotton fiber data is collected from agricultural marketing service. It has daily cotton quality summary data with different attributes like bales, color, leaf, staple, mike, strength, uniformity, trash, bark and grass.

Descriptive statistics for samples:
Retained 42 (70%) samples whose summary statistics are recorded in Table 1. Fiber strength, elongation, trash count and short fiber index, had the highest deviations from CDO’s quality reference values. Coefficients of variation, CV (%) values of the samples, suggest that trash indicators presented the highest variance, followed by short fiber index. This is obvious as these variables are mostly dependent on human and processing factors. Maturity, uniformity index, and whiteness were the least variant. Overall, most parameters were fairly uniform within and inter regions shows in Fig.1
Fig. 1: Descriptive summary of cotton quality measurements from USTER HVI 1000

### Quality Attribute

<table>
<thead>
<tr>
<th>Quality Attribute</th>
<th>abbr</th>
<th>unit</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>stdev</th>
<th>CV%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color grade: yellowness</td>
<td>Rd</td>
<td>%</td>
<td>9.7</td>
<td>12.4</td>
<td>10.5</td>
<td>0.8</td>
<td>7.6</td>
</tr>
<tr>
<td>Color grade: whiteness</td>
<td>b</td>
<td>%</td>
<td>73</td>
<td>76.2</td>
<td>74.4</td>
<td>0.79</td>
<td>1.1</td>
</tr>
<tr>
<td>Leaf grade</td>
<td>Lfgd</td>
<td>%</td>
<td>0.1</td>
<td>0.7</td>
<td>3</td>
<td>2</td>
<td>35.2</td>
</tr>
<tr>
<td>Short fiber index</td>
<td>SFI</td>
<td>%</td>
<td>6.1</td>
<td>8.8</td>
<td>7.2</td>
<td>0.77</td>
<td>80.7</td>
</tr>
<tr>
<td>Uniformity index</td>
<td>UI</td>
<td>%</td>
<td>83</td>
<td>85.9</td>
<td>84.2</td>
<td>0.78</td>
<td>0.9</td>
</tr>
<tr>
<td>Fiber strength</td>
<td>Str</td>
<td>%</td>
<td>26.6</td>
<td>30.4</td>
<td>28.2</td>
<td>0.95</td>
<td>3.4</td>
</tr>
<tr>
<td>Maturity</td>
<td>Mat</td>
<td>%</td>
<td>0.56</td>
<td>0.88</td>
<td>0.87</td>
<td>0.007</td>
<td>0.5</td>
</tr>
<tr>
<td>Fiber Elongation</td>
<td>Elg</td>
<td>%</td>
<td>43</td>
<td>62</td>
<td>53</td>
<td>0.44</td>
<td>83</td>
</tr>
</tbody>
</table>

**A. Accuracy:**

Social Accuracy is defined as the proportion of true positives and true negatives among the total number of results obtained. Accuracy is evaluated as,

\[
\text{Accuracy} = \frac{\text{True positive} + \text{true negative}}{\text{True positive} + \text{true negative} + \text{False positive} + \text{False negative}}
\]

Fig.1 shows the comparison of accuracy between existing method and proposed method for classification of quality of cotton fiber.

![Comparison of Accuracy](image)

**Fig. 2: Comparison of Accuracy**

Fig.2 shows the comparison of accuracy between existing PCA, AHC,K-means based classification and proposed PCA and FCM based classification in distributed platform. X axis represents the number of attributes and Y axis represents the accuracy value. From the fig.1 it is proved that the proposed classification method has high accuracy than the existing method.

**B. Precision:**

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

\[
\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}
\]

Fig.2 shows the comparison of precision between existing method and proposed method for classification of quality of cotton fiber.
Fig. 3: Comparison of Precision

Fig. 3 shows the comparison of precision between existing PCA, AHC, K-means based classification and proposed PCA and FCM based classification in distributed platform. X axis represents the number of attributes and Y axis represents the precision value. From the fig.1 it is proved that the proposed classification method has high precision than the existing method.

C. Recall:

The Recall value is evaluated according to the classification of data at true positive prediction, false negative.

\[
\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}
\]

Fig. 2 shows the comparison of recall between existing method and proposed method for classification of quality of cotton fiber.

Fig. 4: Comparison of Recall

Fig. 4 shows the comparison of recall between existing PCA, AHC, K-means based classification and proposed PCA and FCM based classification in distributed platform. X axis represents the number of attributes and Y axis represents the recall value. From the fig.1 it is proved that the proposed classification method has high recall than the existing method.

D. Time Complexity:

Time complexity is a measure of amount of time taken by the methods. Fig. 4 shows the comparison of time between existing method and proposed method for classification of quality of cotton fiber.
Fig. 5: Comparison of Time

Fig. 5 shows the comparison of time between existing PCA, AHC, K-means based classification and proposed PCA and FCM based classification in distributed platform. X axis represents the methods and Y axis represents the time value. From the fig. 2 it is proved that the proposed classification method has less time complexity than the existing method.

Conclusion:

The classification of quality of cotton fibre is performed in distributed environment. Initially the cotton fibre data are collected from agricultural marketing service. Then the features are extracted using PCA in distributed platform. The extracted features are used in FCM clustering. These are implemented in distributed environment. The experimental results prove that the proposed methods have high accuracy, precision, recall and less time complexity than the existing methods.

REFERENCES